Les dangers de la loi renseignement : « Des dizaines de milliers de personnes vont être suspectées à tort »

jeudi 14 mai 2015, par CASTELLUCCIA Claude, LE METAYER Daniel, UNTERSINGER Martin (Date de rédaction antérieure : 6 mai 2015).

Le projet de loi sur le renseignement a été adopté, mardi 5 mai, en première lecture par l'Assemblée nationale.

L'un des dispositifs les plus critiqués par les nombreux acteurs de la société civile opposés à ce texte est celui des « boîtes noires » : ces dispositifs seront installés chez les fournisseurs d'accès à Internet et ingéreront de grandes quantités de données. Un algorithme détectera ensuite de potentiels comportements terroristes sur Internet.

Daniel Le Métayer et Claude Castelluccia sont tous deux directeurs de recherche à l'Institut national de recherche en informatique et en automatique (Inria), spécialistes en protection de la vie privée. Ils alertent, comme d'autres avant eux notamment sur le site Rue89, sur les dérives possibles d'un tel dispositif ainsi que sur son inefficacité pour lutter contre le terrorisme.

Martin Untersinger : La loi prévoit que les données traitées par l'algorithme présent dans les boîtes noires soient anonymes ; cela vous rassure-t-il ?

Daniel Le Métayer : Non, ça ne me rassure pas. D'un point de vue technique, l'anonymat des données signifie qu'on ne peut pas retrouver l'identité de la personne. Or, ici, tout est prévu pour la retrouver si elle est suspecte. Ce ne sont pas des données anonymes. On utilise le mot anonyme pour rassurer, on joue sur les mots.

Claude Castelluccia : On ne sait pas anonymiser des métadonnées. Si on le faisait, ces dernières seraient agrégées ou modifiées, et donc inutilisables dans ce contexte (identification de suspects). Par ailleurs, l'historique des connexions sur Internet est unique, et constitue un identifiant qui permet parfaitement de réidentifier une personne.

Dans les boîtes noires figurera un algorithme censé détecter des comportements terroristes. Cela vous paraît-il réaliste ?

CC : Tout algorithme peut produire ce qu'on appelle des « faux positifs », c'est-à-dire des résultats qui ne correspondent pas à ce qu'il cherche en réalité. Dans ce cas précis, comme le nombre d'individus qu'on cherche à détecter est marginal, parce que heureusement il y a très peu de terroristes par rapport à la taille de la population, on va identifier de nombreuses personnes innocentes. C'est inévitable, même avec des algorithmes très performants ! C'est ce que les mathématiciens appellent le « paradoxe des faux positifs ».

DLM : Même avec un algorithme extrêmement précis, qui ne se tromperait qu'une fois sur cent, à l'échelle de la population il y aurait de l'ordre de 600 000 personnes suspectées à tort. 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 d'environ 0,5 %, ce qui est négligeable. Cela a des conséquences du point de vue des droits de l'homme, mais aussi de la stricte effectivité : surveiller des dizaines ou des centaines de milliers de personnes, ce n'est pas tenable et pas très rationnel.

CC : De plus, pour être efficace, l'algorithme aura besoin d'un profil de terroriste précis, ce qui n'est pas gagné : il n'est pas évident qu'il existe un profil précis et unique pour tous les terroristes. De plus, même si c'était le cas, le nombre de terroristes étant limité, les techniques du « big data », qui permettent de définir des modèles précis à partir d'une quantité considérable de données, ne sont pas directement applicables ici.

La Commission nationale de contrôle des techniques de renseignement (CNCTR) comprendra un expert, nommé par l'Arcep, l'Autorité des télécommunications, qui pourra expertiser l'algorithme, est-ce suffisant ?

DLM : Comment un seul expert pourrait faire face à ce genre de situation ?!

CC : Les techniques et les compétences à posséder (réseaux, bases de données, statistique, cryptographie, sécurité…) sont tellement complexes et nombreuses qu'il faut toute une équipe. Une personne ne suffit pas.

Est-ce une bonne chose que l'algorithme ne s'intéresse qu'aux seules données de connexion, c'est-à-dire aux métadonnées, et pas au contenu des communications ?

CC : Cela ne change rien. A partir des métadonnées on peut identifier une personne et apprendre beaucoup d'information sur un individu.

DLM : Encore une fois, le terme « métadonnée » est utilisé de manière trompeuse, pour rassurer, en donnant l'impression que la collecte des métadonnées est anodine, alors qu'elles 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 elle-même, qui peut se réduire à une simple prise de rendez-vous.

CC : Un autre exemple est celui des données de géolocalisation, qui sont aussi des métadonnées. Elles permettent par exemple de deviner les lieux de domicile ou de travail, et même la religion d'une personne (qui est une donnée sensible au sens de la loi informatique et liberté), si elle se rend, par exemple, tous les dimanches à l'église. Les sites visités en disent beaucoup sur une personne. Il est faux de dire qu'il est anodin ou pas dangereux de collecter des métadonnées.

Le gouvernement s'est défendu à plusieurs reprises de stocker massivement des données, et le premier ministre a assuré dans une lettre à un député qu'aucune (méta)donnée ne serait stockée. Autrement dit, l'algorithme travaillerait « à la volée » : est-ce possible ?

CC : Tout dépend des algorithmes utilisés. Certains travaillent sur des séquences d'événements qui durent dans le temps, et dans ce cas il me paraît difficile de travailler à la volée. Imaginez un algorithme fictif très simple qui identifie un suspect s'il visite, par exemple, au moins 5 sites différents parmi une liste 1 000 sites « suspects ». Dans ce cas, comment savoir si un individu a visité au moins 5 sites différents, si on n'a pas sauvegardé les 4 sites suspects qu'il a visités préalablement ? Donc, sauf à utiliser des algorithmes de classification très basiques, et probablement peu performants, je doute qu'on puisse obtenir des performances acceptables sans stocker aucune métadonnée.

Que faudrait-il pour introduire des garde-fous techniques dans cette loi ?

DLM : La notion de « responsabilité », dans le sens de l'« accountability » des Anglo-Saxons, le fait de devoir rendre des comptes, est une condition sine qua non de la confiance, et une contrepartie nécessaire de tout pouvoir est primordiale. Pour cela, les mesures techniques de contrôle doivent assurer que les données collectées sont authentiques et intègres, et qu'elles ne peuvent être consultées que par les agents habilités et authentifiés, pour des finalités bien définies et avec une traçabilité des autorisations. Sans ces garanties techniques, tout contrôle par la CNCTR risque de se révéler illusoire.

CC : On s'est focalisé ce dernier mois sur la surveillance gouvernementale, on en a un peu oublié la surveillance commerciale, où il y a un manque de transparence total. Il serait intéressant d'avoir ce débat, d'autant plus que l'argument de la surveillance commerciale a été utilisé par les défenseurs du projet de loi. Les entreprises sont à la pointe du profilage, elles sont souvent situées à l'étranger, et dépensent beaucoup d'argent dans ce domaine. C'est, à mon avis, un vrai danger, voire un plus grand danger.

Cet entretien reflète exclusivement l'opinion de ses auteurs et n'engage en aucune façon l'Inria.

Propos recueillis par Martin Untersinger
Journaliste au *Monde*

P.-S.

* « Loi renseignement : « Des dizaines de milliers de personnes vont être suspectées à tort » ». Le Monde.fr | 06.05.2015 à 09h38 • Mis à jour le 06.05.2015 à 14h55.