La Jaune et la Rouge mars 2014

Big Data

Les trois paradoxes de la sécurité

"Cyberwar is storytelling", Martin C. Libicki

Par Philippe Wolf (78), ingénieur général de l'armement, ANSSI

De récentes cyber-attaques ont révélé les dangers liés à l'existence de données de masse. Ces dangers présentent des caractéristiques propres qu'il est important de bien identifier pour adapter la politique de sécurité des systèmes d'information (SSI).

REPERES

Le terme *big data* fait référence à des ensembles de données dont la taille dépasse la capacité des logiciels usuels pour collecter, gérer et traiter les données dans un temps raisonnable. Les traitements de masse impliquent une nouvelle approche de la donnée : collecter et utiliser beaucoup de données plutôt que de se contenter d'échantillons comme l'ont fait des générations de statisticiens ; accepter de traiter des données imparfaites ou mal organisées, une part d'inexactitude peut en général être tolérée car dans de nombreux cas, il est plus avantageux d'avoir bien plus de données qu'un plus petit nombre de données très exactes ou finement sélectionnées afin d'être représentatives ; accepter de renoncer à rechercher des causalités au profit de la recherche de corrélations, de motifs qui peuvent aider à prédire le futur ; le *big data* aide à répondre à la question du quoi et pas à celle du comment, ce qui est souvent suffisant.

Les problèmes de sécurité liés au *big data* sont multiformes suivant l'origine des données (publiques, privées ou mixtes), la loyauté de leur recueil, la présence ou non, directe ou indirecte, de données personnelles, l'objectif poursuivi (bien commun scientifique ou avantage concurrentiel), la transparence ou l'opacité des buts poursuivis, les infrastructures (publiques, privées ou mixtes) de stockage et de calculs mises en œuvre et le caractère ouvert ou fermé des traitements algorithmiques. Les attaques possibles contre le *big data* sont, de ce fait, multiples : attaques informatiques classiques, atteintes aux infrastructures, usages détournés des puissances de calculs, mais aussi, clonages de masses frauduleux, falsifications parfois partielles des données, manipulations de l'information ou encore atteintes aux personnes dans leur dignité.

Diagnostic

Posons d'abord un diagnostic qui s'appuie sur trois paradoxes de la finalité du *big data*, soulignés par un juriste et un stratège du "Cloud"[1], que nous résumons (en italiques) et commentons. Une difficulté apparaît qui est de ne pas analyser cette nouvelle manière d'acquérir des connaissances en ne considérant que ceux qui ont actuellement la capacité de collecter et d'exploiter des données en masse

[1] Neil M. Richards & Jonathan H. King, *THREE PARADOXES OF BIG DATA*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325537

à des fins commerciales (Google, Facebook, Twitter, …) ou d'espionnage (NSA et autres services de renseignements).

***En exergue* : Les problèmes de sécurité liés au *big data* sont multiformes**

Le paradoxe de la transparence

« La "privacy" ou « protection des données personnelles » est morte ; nos données personnelles deviennent transparentes. Les traitements big data devraient aussi l'être ; mais ce sont les « écosystèmes numériques fermés » qui les manipulent avec le secret le plus absolu. Les décisions prises par les robots de surveillance sont d'une opacité kafkaïenne. »

Le modèle économique « gratuit » de l'Internet repose sur une nouvelle forme de commercialisation et de valorisation de données collectées sur le comportement des personnes. « Quand vous ne voyez pas le service [payant], c'est que vous êtes le produit [revendu] ! »[2]. Il se nourrit de l'une de deux visions irréconciliables du droit à un espace privé[3] que soulignent les difficiles débats européens sur la protection des données personnelles qui serait, vue d'Amérique du Nord, un obstacle à l'innovation[4]. Les objectifs de cette protection sont le respect des personnes dans les traitements par le recueil de leur consentement préalable, le contrôle de leurs finalités, la limitation de la collecte[5] et des croisements. Mais la diffusion des ordiphones, par exemple, pousse à la personnalisation de l'action sur les informations ; elle accentue le décalage entre les besoins de services publics (éducation, santé, régulation économique, ordre public) et la tentative de privatisation des données directement liées aux personnes solvables (pour la publicité directe).

Connaître tout sur ses clients

Parmi les géants de la Toile (Web), tous américains, la société Google s'appuie sur la recherche « en psychologie cognitive » pour mieux atteindre son but d' « amener les gens à utiliser leur ordinateur avec plus d'efficacité » ; elle ne sera pas satisfaite tant qu'elle ne disposera pas de « 100 % des données de ses utilisateurs »[6]. Elle utiliserait même la biométrie de la frappe clavier pour reconnaître l'usager derrière sa machine.

[2] Ari Melber, *The Secret to Facebook's IPO Value*, http://www.thenation.com/blog/166388/secret-facebooks-ipo-value

[3] James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 4 janvier 2004, http://www.yalelawjournal.org/images/pdfs/246.pdf

[4] Pour ne citer qu'elle, la société française CRITEO, cotée au Nasdaq depuis peu, vend des services robotisés produisant en quasi temps réel, les bandeaux publicitaires ciblés en fonction des données identifiant l'internaute.

[5] "a priori data minimization", Datasparsamkeit (voir les règles pour le casier et les archives publiques judiciaires en France).

[6] *Internet rend-il bête ?*, Nicholas Carr, Robert Laffont, 6 octobre 2011.

Le paradoxe de l'identité

« Le droit à l'identité, au moi, à l'ego nécessite le libre arbitre. Les robot-programmes béhavioristes du big data cherchent à identifier qui nous devons être, qui nous devons aimer, ce que nous devons consommer, ce qui nous est interdit. Jusqu'à influencer nos choix intellectuels et nous faire perdre notre identité. »

Le film *Bienvenue à Gattaca* anticipait cette uniformité engendrée par des machines. Ce n'est encore, heureusement, que de la science-fiction. Par contre, l'hyper-connectivité accroît, sans pause, notre dépendance cybernétique. La publicité d'une marque allemande de voitures « haut de gamme » vante son attrait irrésistible : « CONNECTED DRIVE. Mieux connecté. Encore plus libre »[7].

Les réseaux sociaux sont l'archétype d'une illusion numérique généralisée. Pour le meilleur, comme la gestion en temps réel de catastrophes naturelles, la diffusion instantanée de l'état du monde ou la sortie de l'isolement qu'engendrent nos sociétés trop individualistes. Mais aussi pour le pire, comme l'exploitation outrancière de leurs capacités d'intrusion dans les intimités.

Le droit à l'oubli, une chimère par nature, devient une demande à satisfaire par la limitation de la collecte. Il ne fonctionne que pour nos données domestiques par la fragilité, souvent ignorée, des supports numériques personnels. Alors, qu'à l'image du sparadrap du capitaine Haddock, ce que nous aimerions voir disparaître dans les volutes du passé risque fort de rester dans l'éponge Internet.

Le cas NSA

Dans les révélations Snowden, on apprend ainsi que le programme SYNAPSE de la NSA vise à stocker, pour chaque internaute, 94 critères d'identité (n° téléphone, e-mail, adresses IP, etc.) permettant d'y corréler 164 types de relations (profilage par les réseaux sociaux, paiements électroniques, profils d'intérêts, déplacements grâce à la géolocalisation, etc.)[8]. On est très loin du principe de non croisement des données et du respect d'un espace privé prévu dans la *Loi Informatique et Libertés*. Mais, le diable avance masqué et toujours dans la séduction.

***En exergue* : Les réseaux sociaux sont l'archétype d'une illusion numérique généralisée**

Le paradoxe du pouvoir

« Le big data est censé nous fournir une boîte à outils pour mieux comprendre le monde. Mais ses robots sont entre les mains d'institutions intermédiaires, qui ont le pouvoir de manipulation, et non des individus. Le big data créera des vainqueurs et des vaincus[9]*. »*

Noam Chomsky, dans une conférence récente, constate que le pouvoir lié à la possession des données existe depuis une centaine d'années mais que la surprise vient aujourd'hui des échelles atteintes. Il

[7] Voir http://www.bmw.fr/fr/topics/innovation/connecteddrive-2013/overview.html
[8] Voir http://mobile.nytimes.com/2013/09/29/us/nsa-examines-social-networks-of-us-citizens.html
[9] Louis Pouzin, *Où va l'internet ? Mondialisation et balkanisation*, http://www.eurolinc.eu/spip.php?article79

rappelle aussi que le « pouvoir demeure fort quand il reste dans le noir ; exposé à la lumière du soleil, il commence à s'évaporer » [10].

La révolution scientifique promise par le *big data*[11], permettrait l'élaboration de nouvelles théories scientifiques libérées des capacités « réduites » du cerveau humain qui migreraient du conceptuel déductif vers l'inductif ; même si l'intuition humaine et quelques résultats théoriques (théorèmes d'incomplétude de Gödel) devraient encore éloigner pour un temps le spectre d'une intelligence artificielle dominatrice. Il faudra, pour la communauté scientifique mondiale, plus de banques de données ouvertes. Mais, depuis 2010, les banques publiques génomiques ne sont plus exhaustives, pour des raisons budgétaires, marquant ainsi un retour vers la marchandisation du vivant.

Coûts prohibitifs

L'utopie de la bibliothèque mondiale de tous les savoirs, chère à J.L. Borgès[12], s'éloigne devant les coûts des centres de traitements énergivores. Pourtant le progrès de l'humanité passe par une coordination négociée, décentralisée, multilinguistique et multiculturelle dans l'acquisition et la maîtrise des savoirs, des biens mondiaux.

Les traitements eux-mêmes relèveront parfois du logiciel libre et ouvert (astronomie, génomique, recensement de la faune et de la flore, pharmacologie, démographie, physique des particules, météorologie, climatologie, macro-économie, sociologie) mais bien plus souvent de solutions propriétaires, au nom de la protection du patrimoine informationnel et du secret des affaires des grandes entreprises. Il ne s'agit pas de les opposer mais d'imaginer les mécanismes sécurisés créant les passerelles nécessaires. Il faudra également, sans naïveté ni excès, parfois limiter la capacité de ces nouveaux "little brothers".

Nouveaux dangers

Un diagnostic étant posé sur le traitement de données en masse, il convient de tenter d'en recenser les dangers, à l'expérience de cyber-attaques récentes.

La visibilité acquise de l'exploitation systématique de vulnérabilités non corrigées, dites 0-day[13], permettant des attaques ciblées surprises, modifie la pratique de la protection en SSI. Les protections périmétriques (le mythe des lignes Maginot et Siegfried) et la surveillance interne des traces ou des comportements sont nécessaires mais ne suffisent plus. La virtualisation et l'ubiquité, constitutives des architectures massives, augmentent les surfaces d'attaques et les délocalisent. Les efforts et les budgets de sécurisation doivent alors se concentrer sur les données les plus sensibles. Le nomadisme condamne, de toutes les façons, les autres données à une transparence forcée.

Une sécurité « à la volée »

[10] Voir http://www.hyperorg.com/blogger/2013/11/15/liveblog-noam-chomsky-at-engaging-data/
[11] Ce que Jim Gray appelle le "fourth paradigm", voir http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_lynch.pdf
[12] Voir http://fr.wikipedia.org/wiki/Tl%C3%B6n,_Uqbar,_Orbis_Tertius
[13] *Vulnérabilités 0-Day, prévention et bonnes pratiques*, http://www.ssi.gouv.fr/fr/bonnes-pratiques/recommandations-et-guides/securite-du-poste-de-travail-et-des-serveurs/vulnerabilites-0-day-prevention-et-bonnes-pratiques.html

Les modèles de sécurité statiques qui protègent nos systèmes ont une quarantaine d'années (Multics, Unix). Il y a urgence à les repenser autour de concepts de dynamique et de proactivité. Le *big data* en mode flux (streaming) oblige à gérer une sécurité « à la volée ». L'irruption de l'internet (une architecture faible des années 1970) comme système d'information global dès 1991 n'avait pas anticipé les enjeux de protection. La plasticité des protocoles non sécurisés de l'Internet devra pourtant s'accommoder de nouveaux services sécurisés, notamment pour la protection des données personnelles mais aussi patrimoniales.

***En exergue* : la malveillance d'un code informatique est indécidable**

Ces services sécurisés seront bâtis à partir de briques cryptographiques incontournables, mais devront porter une attention plus grande à la facilité d'emploi[14]. Même si leur usage ne pourra jamais être transparent et se passer de la gestion humaine (comme certains voudraient nous le faire croire en remplaçant les administrateurs par des automates !). Il faudra enfin considérer que la protection des données et informations (accessibilité, authenticité, contrôle des finalités) au moyen de la cryptographie (attaches indélébiles de marques, de signatures, obscurcissement) est un moyen faillible, au-delà des mathématiques « parfaites » sous-jacentes.

Enfin, les quatre V (Volume, Variété, Vélocité, Véracité) associés au *big data* obéissent aux limitations de deux théorèmes démontrés en 2002[15]. Ceux-ci sont à rapprocher du théorème du virus de 1986[16] qui dit que la malveillance d'un code informatique est indécidable. Ces incertitudes inhérentes au *big data* changent la donne en matière de défense et de sécurité des systèmes d'information.

Dualité

Il faut rappeler que toute fonction de sécurité est à usage dual ; elle servira aussi bien le criminel que l'honnête homme. Cela ne doit pas justifier le piégeage généralisé (matériel, logiciel ou sémantique et mathématique) qui pénalise, avant tout, la cyber-protection.

[14] Alma Whitten et J. Doug Tygar, *Why Johnny can't encrypt ? a user experiment of PGP 5.0*, http://www.cs.cmu.edu/~alma/johnny.pdf. Alma Whitten travaille maintenant chez Google.

[15] Le théorème de Brewer qui affirme qu'il est impossible de satisfaire à la fois la cohérence, la disponibilité et la résistance au morcellement (voir http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.1495&rep=rep1&type=pdf) et le théorème de Kleinberg identique au précédent pour la cohérence, la disponibilité et l'invariance d'échelle (voir http://www.cs.cornell.edu/home/kleinber/nips15.pdf).

[16] Fred Cohen, *Computer Viruses*, Janvier 1986, http://all.net/books/Dissertation.pdf

La Jaune et la Rouge mars 2014

Big Data

Garantir la disponibilité, l'intégrité et la confidentialité des données

Par Philippe Wolf (78), ingénieur général de l'armement, ANSSI

Les trois fonctions principales de la sécurité des systèmes d'information sont la disponibilité, l'intégrité et la confidentialité. Une analyse des remèdes autour des trois aspects complémentaires que sont la sécurité des infrastructures, la protection des données et la protection des informations produites, ouvre de nouvelles pistes de recherche ou de développement.

REPERES
La CNIL propose que « l'appellation " *coffre-fort numérique* ", ou " *coffre-fort électronique* ", soit réservée à une forme spécifique d'espace de stockage numérique, dont l'accès est limité à son seul utilisateur et aux personnes physiques spécialement mandatées par ce dernier. Les services de coffre-fort numérique doivent garantir l'intégrité, la disponibilité et la confidentialité des données stockées et impliquer la mise en œuvre des mesures de sécurité décrites dans la recommandation ».

Sécurité des infrastructures

La sécurité des infrastructures mettant en œuvre le *big data*, potentiellement sensibles au regard de ce qu'elles manipulent, fait appel principalement aux fonctions de disponibilité et d'intégrité. La résilience doit être une propriété globale de la chaîne (réseaux, baies, procédures, humains) et ne peut s'appréhender, dans son ensemble, qu'avec une analyse holistique et une gestion permanente des risques. Du très classique, même si les questions d'interdépendance et d'éparpillement prennent une importance cruciale due à la complexification des architectures de protection. De plus, le modèle du pair-à-pair se substitue au modèle client-serveur qui facilitait la supervision de sécurité. L'introduction de mécanismes de sécurité sur des couches logicielles qui se standardisent (openstack, hadoop[17], etc.) doit pouvoir apporter une résistance nouvelle.

Les puissances de calcul requises par le *big data* nécessitent, sauf pour quelques très grosses entreprises, d'externaliser ou, au minimum, de mutualiser stockages et traitements dans l'infonuagique (cloud computing). Le recours au *cloud computing* réclame des précautions même dans le cas de transparence absolue.

Quatre règles à respecter

Les quatre trayons du « cloud maitrisé ou souverain » sont connus mais pas toujours activés : faire appel à un ou des prestataire(s) de confiance ; être capable d'auditer réellement la solution dans un temps court ; avoir la garantie testée de réversibilité pour changer de prestataire, sans pertes, si nécessaire ; rédiger les contrats sous la protection du droit national pour gérer le risque juridique.

[17] En particulier, sa surcouche Accumulo qui propose du contrôle d'accès par marquage (sans chiffrement), premier pas vers le contrôle de finalité de la collecte des données. Deux sociétés au moins supportent ce logiciel libre, voir http://accumulo.appache.org

***En exergue* : rédiger les contrats sous la protection du droit national pour gérer le risque juridique**

Protection des données

Dans le cas du *big data* non ouvert (privé), la confidentialité des données stockées ne pose pas de problèmes particuliers si l'entreprise ou l'organisme garde la capacité de gérer ses propres clés de chiffrement ou de signature, de préférence dans un coffre-fort numérique labellisé[18], ou en confie la gestion à des tiers réellement de confiance. Pour rendre confidentiels les algorithmes de calcul, il manque aujourd'hui un ingrédient essentiel qui serait une implémentation pratique du chiffrement dit homomorphique , c'est-à-dire d'un chiffrement qui donnerait un moyen de réaliser diverses opérations sur le chiffré sans recourir à l'opération de déchiffrement complète. Une avancée dans ce domaine comme sur le calculateur quantique ou à ADN nécessitera, de toutes les manières, de reconcevoir une algorithmique adaptée.

L'intégrité classique qui repose sur la signature numérique doit être, à son tour, révisée. Il existe déjà des dérives potentielles liées aux calculs largement répartis ou en grilles. Le respect des règles internationales de non-prolifération impose un contrôle, préalable de préférence, à un usage dévoyé des puissances calculatoires disponibles. La seule signature des ressources partagées, distribuées, hétérogènes, délocalisées et autonomes ne suffit plus. Des techniques d'obscurcissement (« obfuscation de code ») compliquent le contrôle.

Tolérance au flou

L'intégrité stricte des données n'est plus nécessaire quand il s'agit de manipuler des données non structurées, parfois faussées ou incomplètes ou de travailler principalement par échantillonnage. Une tolérance au flou, aux calculs approchés et aux mutations rompant le clonage binaire parfait, sont des ingrédients porteurs d'une meilleure adéquation du *big data* au monde réel qu'il est censé nous aider à comprendre. Cette nouvelle intégrité devrait vérifier les divers paramètres d'une donnée : attribut et granularité fixée initialement ; accessibilité ; authenticité ; contrôle des finalités dont la dissémination.

Risques d'identification

Un pan croissant du *big data* touche aux données personnelles quand elles n'en sont pas le carburant premier[19]. Les progrès des moteurs de recherche intelligents permettent facilement d'identifier une personne à partir d'un nombre très réduit de caractères, cela d'autant plus que l'intimité est littéralement mise à nue dans les réseaux sociaux. On retrouve à une échelle nouvelle, de vieux problèmes d'inférences par déduction, induction, abduction ou adduction dans les bases de données classiques. Les croisements de données permettent des attaques par canaux auxiliaires sémantiques -

[18] Voir http://www.cnil.fr/linstitution/actualite/article/article/adoption-dune-recommandation-sur-les-coffre-forts-electroniques/

[19] Lire, de ce point de vue, la dernière recommandation ENISA sur la protection de la vie privée dans les données de connexions, http://www.enisa.europa.eu/media/news-items/enisa-publishes-new-study-for-securing-personal-data-in-the-context-of-data-retention

attaques qui ne visent pas directement les protections théoriques mais leur implémentation pratique-structure redoutée en SSI. On arrivait à négliger ou à juguler les canaux cachés numériques : ce n'est plus le cas avec les canaux sémantiques[20].

Quatre critères sécuritaires

Les critères communs pour l'évaluation de la sécurité des technologies de l'information[21] introduisent dès 1999 sous l'impulsion du Dr Pfitzmann des fonctions de sécurité pour la protection des données personnelles. Elles sont au nombre de quatre. L'anonymat garantit qu'un sujet peut utiliser une ressource ou un service sans révéler son identité d'utilisateur. La possibilité d'agir sous un pseudonyme garantit qu'un utilisateur peut utiliser une ressource ou un service sans révéler son identité, mais peut quand même avoir à répondre de cette utilisation. L'impossibilité d'établir un lien garantit qu'un utilisateur peut utiliser plusieurs fois des ressources ou des services sans que d'autres soient capables d'établir un lien entre ces utilisations. La non-observabilité garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres, en particulier des tierces parties, soient capables d'observer que la ressource ou le service est en cours d'utilisation.

Ces fonctions font l'objet de travaux algorithmiques novateurs (assainissements des données, k-anonymat, l-diversité, algorithme de Mondrian, calcul multipartite sécurisé, etc.), principalement en Europe, mais tardent à s'implanter dans les traitements numériques de masses qui vont passer rapidement aux traitements d'informations en masses.

Anonymat et santé

La sphère santé-social accumule les difficultés malgré les promesses du *big data* (études épidémiologiques, dossier médical personnel, optimisation des systèmes sociaux). Le constat de départ est qu'il n'y a pas de confidence (médicale) sans confiance (singulière). Il faut alors distinguer la confidentialité-discrétion partageable par du chiffrement réversible de la « confidentialité-séclusion[22] » qui nécessite des fonctions à sens unique. Mais dans ce dernier cas, la pseudo-anonymisation réversible serait parfois préférable à une véritable anonymisation irréversible, dans le cas, par exemple, de détection d'une maladie orpheline ou d'une grave épidémie où il faudrait retrouver l'individu porteur. Il manque clairement un modèle de sécurité partagé.

Protection des informations

On ne peut éliminer le rôle du sujet dans la production de l'information, ou parfois de la connaissance, par le *big data*. « La signification d'une information est toujours relative »[23]. Il s'agit de mesurer l'intelligibilité, la vérifiabilité et la traçabilité, d'estimer la responsabilité contractuelle, de gérer les conflits d'influences, de distinguer les fausses nouvelles, bref de résister au mirage du *big data* simpliste.

[20] Comme, par exemple, le piégeage mathématique de la norme de génération de nombres pseudo-aléatoires Dual_EC_DRBG par la NSA, publiée en 2006 par le NIST et dont la porte dérobée n'a été découverte qu'en 2008.
[21] Voir http://www.ssi.gouv.fr/fr/certification-qualification/cc/les-criteres-et-methodologies-d-evaluation.html
[22] Termes introduits par Gilles Trouessin, https://www.ossir.org/jssi/jssi2008/4B.pdf
[23] Jean Zin, *Le monde de l'information*, 2004, http://jeanzin.fr/ecorevo/sciences/mondinfo/mondinfo.htm

En exergue : **Des amendes record touchent aujourd'hui des institutions financières**

Des amendes record touchent aujourd'hui des institutions financières. Elles sanctionnent des infractions à répétition (subprimes, Libor, Euribor, taux de changes, marché pétrolier) qui n'auraient pas été possibles sans *l'obscurcissement numérique* , technique consistant à cacher des informations en les noyant dans une masse de données (High Frequency Trading, « flash crashes », complexification des produits financiers, l'argent est un produit). L'infobésité, sans diète, nourrit et amplifie cette obscurité. De plus, les biais cognitifs du *big data*, voulus ou non, aveuglent une saine compréhension des enjeux de sécurité.

La capacité d'absorption humaine étant limitée[24], un différentiel de plus en plus grand se créera avec les capacités attendues des robots-programmes. Quand les résultats espérés ne seront pas là, la tendance sera de complexifier les traitements par une massification encore plus grande des données et par l'ajout de paramètres aux automates. Alors qu'il faudrait, au contraire, modéliser, analyser, expliquer et mieux cibler et cribler les données utiles et rationaliser cette intelligence artificielle (« diviser pour traiter », par exemple). Cette tendance à l'entropie porte en elle le germe des « *accidents de la connaissance* » pointés par l'essayiste Paul Virilio[25]. À brasser trop large et trop gros, on oublie les fonctions essentielles (des exemples actuels : logiciel Louvois, écotaxe, PRISM, Obamacare, fuites de la NSA, etc.) et on bride l'engagement.

Sciences du danger et *big data*

Il est intéressant de noter que les cindyniques, ou sciences du danger, commencent à investiguer le champ de l'information[26]. Elles proposent un regard à cinq dimensions, examinant à la fois : la dimension des données (axe statistique) ; la dimension des modèles (axe épistémique) ; les finalités de l'acteur (axe téléologique) ; l'axe des règles, normes, codes auxquels est soumis (ou que s'impose) l'acteur ; et les valeurs (éthiques, morales) de l'acteur (axe axiologique).

Une nouvelle approche de la SSI

Le *big data* ouvre aussi des perspectives nouvelles en SSI, qui passent d'abord par la mutualisation des compétences devant une menace multiforme qui s'adapte très vite aux mutations technologiques. Dans cette lutte aujourd'hui inégale entre défenseurs et attaquants, l'analyse des signaux faibles est largement prônée. Le *big data* semble adapté à cette détection d'anomalies sur l'échelle dite des sources ouvertes. Il prépare l'analyse des significations (la sémantique) des affrontements cyber. Il fournit un faisceau d'indices permettant aux analystes d'évaluer l'origine des attaques. Il doit aussi servir à anticiper les usages malveillants des technologies micro-robotiques constitutives de l'Internet

[24] Un humain absorbera au plus 40 pétaoctets (10^15) dans sa vie, à rapporter aux 200 yottaoctets (10^24) manipulés par Internet sur 60 ans.
[25] *L'accident originel*, Paul Virilio, Galilée, 2005.
[26] Voir http://ifrei.org/tiki-index.php?page=InfoCindynique.

des objets. Enfin, il doit offrir des simulations dynamiques d'attaques, les plus proches du réel, pour en déduire les mécanismes de contre-réaction les plus pertinents.

Plusieurs écueils constitutifs du *big data* sont à éviter ici. Il ne s'agit ni de remplacer la précision des données par leur masse, ni de remplacer la recherche de causes par celle de coïncidences ou de corrélations. Il faut se méfier du retour de certaines illusions bien connues des informaticiens expérimentés, comme l'apprentissage, les réseaux de neurones, voire certains aspects de l'intelligence artificielle dans lesquelles les hypothèses implicites (structure du réseau de neurones, biais de la collecte servant à l'apprentissage) ne peuvent être ignorées. Appliquées, par exemple, à l'identification de suspects ou de cibles en sécurité civile, cela semble porteur de très graves dangers pour les sociétés.

Mais, la SSI ne se réduit pas, malheureusement, aux architectures de systèmes. L'assemblage de composants sécurisés ne garantit pas la solidité du tout ; au contraire, la complexité facilite le travail de l'attaquant dans la recherche d'un chemin d'attaque. A contrario, la monoculture technologique favorise le contrôle centralisé mais cette facilité fragilise également.

En exergue : Pour faire du *big data* un outil de progrès, il faut en maîtriser les dérives

Protéger la cyber-diversité

Une analogie s'impose. La diversité des espèces est le plus grand rempart immunitaire contre la perte d'un écosystème. Aussi, la cyber-diversité[27], si malmenée par quelques écosystèmes numériques fermés dont aucun n'est européen, reste le constituant principal d'une véritable défense en profondeur.

Éthique du *big data* ?

Un rapport gouvernemental récent[28] affirme qu'il est impératif d'« assurer la sécurité des données ». Pour faire du *big data* un outil de progrès sociétal, par exemple pour les villes intelligentes ou "smart cities" (eau, transports, énergie, commerce électronique), il faut en maîtriser les dérives. On pourrait paraphraser le célèbre « Code is Law » de Lawrence Lessig[29] par « microcode is law in cyberspace ». La France ou l'Europe voudront-t-elles revenir dans le jeu technologique ? Une opportunité se présente avec le probable remplacement du silicium par le carbone (graphène).

Quoi qu'il en soit, des règles d'éthiques sont à poser. La France et la vieille Europe sont héritières des vertus de « Dignité, de Réserve et de Droiture » (Epictète). Puissent-elles engager la maîtrise et la

[27] La cyber diversité est en berne. Pour ne prendre qu'un exemple, seules les entreprises IBM (USA) et TSMC (Taïwan, Chine) détiennent les savoir-faire physico-chimiques des fonderies électroniques « silicium ». Les investissements pour créer un nouveau circuit et ses services associés atteignent la somme de 10 milliards de dollars, autant qu'un aéronef.

[28] *Analyse des big data, Quels usages, quels défis ?*, Commissariat général à la stratégie et à la prospective, novembre 2013, Voir http://www.strategie.gouv.fr/blog/wp-content/uploads/2013/11/2013-11-09-Bigdata-NA008.pdf

[29] Traduction française ici : http://www.framablog.org/index.php/post/2010/05/22/code-is-law-lessig et son livre codev2 est ici : http://codev2.cc/

domestication des robots logiciels du *big data* sur une régulation s'inspirant de ces principes en gardant l'homme au centre des enjeux.

Le Centre de données de l'Utah (Utah Data Center) géré par la NSA sera opérationnel en septembre 2014. En juillet 2013, sur la base des plans de la structure, le magazine Forbes estimait la capacité de stockage entre 3 et 12 exabytes (milliards de gigabytes), à l'aide de 10 000 racks de serveurs. (*source wikipedia*).