CPVP

Commission de la
protection de la **vie privée**

Rapport Big Data

Rapporteur : Frank De Smet
assisté par Cliff Beeckman en
Dieter Verhaeghe

Table des matières

Partie 1 – Introduction....3
A. Qu'est-ce que le big data ?....5
B. La chaîne de valeur ("value chain") de big data....6
B.1. Collecte....6
B.2. Stockage et préparation....8
B.3. Analyse....9
B.4. Utilisation....10
Partie 2 - Principes de protection des données et recommandations concernant les analyses de big data....13
A. Remarques générales - méthodologie....14
B. RGPD....15
C. Applicabilité éventuelle de la législation relative à la protection des données lors de l'application de techniques de pseudonymisation ou d'anonymisation, en combinaison ou non avec l'agrégation de données....20
D. "Données à caractère personnel exactes"....24
D.1. Projections techniques optimales (les plus exactes possibles) quant aux personnes physiques....24
D.2. Association par algorithme vs preuve juridique....28
E. Impact individuel ou collectif pour les droits et libertés des personnes concernées....29
E.1. Impact du big data sur les personnes physiques individuelles....29
E.2. Impact social du big data sur certains groupes sociaux....30
F. Principe de loyauté....34
G. Proportionnalité : nécessité, mode de stockage et principe de traitement de données minimal....35
H. Principe de finalité....40
I. Principe de légalité....43
J. Légitimité / Possibilités de considérer les analyses de big data comme un traitement légitime....43
K. Obligation d'information....45
L. Transparence....45
L.1. Transparence d' "informations clés" dans le cadre d'analyses de big data....47
M. Protection de données à caractère personnel sensibles....50
N. Limitations et mesures adéquates en cas de décisions automatisées....51
O. Droits d'accès et de rectification, droit d'effacement des données....53
P. Droit d'opposition....54
Q. Responsabilité du traitement....54
R. Contrôle des traitements....55
S. Sécurité des données à caractère personnel et violations de données à caractère personnel - approche basée sur le risque en vertu du RGPD....57
Lexique....59
Sources....60

Partie 1
Introduction

D'après IBM (2016), nous générons chaque jour quelques 2,5 quintillions d'octets (ou 2,3 trillions de gigaoctets) de données - à tel point que 90 % des données dans le monde ont été créées au cours des deux dernières années seulement [1] . Ces données proviennent d'un peu partout : d'e-mails, d'informations que nous laissons sur les médias sociaux, d'applications mobiles, de photos et de vidéos numériques, de recherches dans Google ou d'autres moteurs de recherche, de capteurs, de transactions d'achats en ligne et hors ligne, de signaux GPS, de téléphones mobiles, de wearables, etc.

L'apport par des technologies innovantes de nouvelles possibilités de retirer de la valeur de ce tsunami de données disponibles peut être désigné par l'appellation évasive de "big data".

Les progrès en matière d'ICT permettent aujourd'hui de collecter, conserver et analyser d'énormes quantités de données. L'analyse automatisée de ce genre de gros fichiers de données permet non seulement de réaliser un important gain de temps mais aussi - si elle est effectuée consciencieusement - d'aboutir à de meilleures informations et connaissances, sur la base desquelles on pourra ensuite potentiellement prendre des décisions plus précises et réaliser des projections plus fiables. Les applications de big data promettent donc de nombreux avantages potentiels : des contrôles plus ciblés dans la lutte contre la fraude, la cartographie plus précise des flux de personnes, des soins de santé sur mesure ("personalised medicine"), etc.

Mais le big data comporte aussi des risques, notamment en ce qui concerne le respect de la vie privée du citoyen. Lors d'analyses de big data, le risque existe ainsi que celles-ci révèlent des informations sensibles à partir d'informations initialement non sensibles (voir le point M ci-après). Il existe par exemple aussi des risques en ce qui concerne l'impact sur les droits et libertés des personnes concernées (voir les points E, N et O ci-après) et le manque de transparence (voir le point L ci-après).

Ce rapport comporte deux parties. Dans la première partie, nous abordons la question de savoir ce que recouvre exactement le big data. Une bonne compréhension du big data est en effet importante pour en évaluer les conséquences sur le plan de la protection de la vie privée. Néanmoins, la distinction entre une analyse de données plutôt classique et une analyse de big data doit être relativisée dans le cadre des débats relatifs à la protection de la vie privée. La réglementation doit être suffisamment générale, solide et neutre technologiquement (ce qu'elle est d'ailleurs pour une grande partie) pour que lors de tout traitement, les personnes concernées soient suffisamment protégées et qu'un bon équilibre soit trouvé afin de ne pas trop réprimer à la fois les droits et libertés des personnes concernées et des opportunités légitimes. Des garanties suffisantes doivent être offertes, permettant un usage légitime et socialement responsable[2]. Ce dernier aspect fait l'objet de la seconde

[1] IBM : https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. Consulté le 18 mai 2016.

[2] On entend par là l'utilisation éthique et socialement consciente des données à caractère personnel. N'en font pas partie les applications qui ont des conséquences discriminatoires pour certains groupes de population (voir ci-après "distorsion de sélection") et qui ne tiennent pas suffisamment compte des droits et libertés des personnes concernées. Voir les points B, F et I ci-après. Voir aussi le document en néerlandais du Wetenschappelijke Raad voor het Regeringsbeleid (Conseil scientifique de la politique gouvernementale des Pays-Bas), Synopsis van WRR-rapport, http://www.wrr.nl/fileadmin/nl/publicaties/PDF-samenvattingen/Synopsis_R95_Big_Data_in_vrije_en_veilige_samenleving.pdf.

partie qui contient une analyse de la législation et de la réglementation actuelles appliquées au big data.

A. Qu'est-ce que le big data ?

Comme c'est très souvent le cas lors de l'apparition d'un nouveau phénomène, il n'existe pas de consensus sur une définition. Il en va de même pour le "big data". Les définitions divergent et mettent respectivement l'accent sur la quantité de données, la diversité et la complexité des données, les nouvelles méthodes d'utilisation des données ainsi que les possibilités sociales, économiques et stratégiques que génère l'utilisation du big data.[3]

Autrement dit, il est difficile de donner une définition univoque du big data. Une meilleure manière de décrire le phénomène ambigu que constitue le big data consiste à relever une série de caractéristiques majeures. Il convient tout d'abord d'observer les propriétés des données utilisées dans le contexte du big data. Elles sont généralement décrites au moyen des 3 lettres "V": *Volume*, *Variété* et *Vitesse*. Dans le contexte du big data, il est généralement question de grandes quantités de données qui doivent être traitées (*Volume*). En outre, ces données proviennent généralement de différentes sources et sont souvent stockées dans des formats structurés, semi-structurés ou non structurés (entre autres texte, image et son) (*Variété*). Dans le cadre du big data, il est aussi question de la vitesse à laquelle les données sont collectées et traitées : des flux de données doivent souvent être traités en continu, parfois en temps réel (*Vitesse*). Au fur et à mesure que le domaine du big data gagnait en maturité, ces 3 "V" ont été complétés par quatre autres "V", parmi lesquels *Véracité* [4], *Valeur* [5], *Visualisation* [6] *et Variabilité* [7].

Ce ne sont toutefois pas uniquement les caractéristiques des données qui déterminent ce qu'est le big data, mais aussi les méthodes ou les techniques d'utilisation de ces données.[8] Pour l'exprimer de manière un peu simple, on peut dire que la méthode scientifique classique part d'hypothèses, concernant généralement des mécanismes causaux, qui sont ensuite testées de manière empirique à l'aide de données rigoureusement collectées à cette fin. Par contre, les méthodes de recherche big data partent généralement de données collectées de manière moins sélective dans lesquelles elles recherchent des modèles ou des corrélations (ce qui n'implique pas nécessairement un lien causal)[9], sans hypothèses préétablies. Alors que l'on pourrait dire de la méthode

---

[3] WRR-rapport 95 (2016). *Big Data in een vrije en veilige samenleving*, pg. 33.

[4] Le taux de précision des données, car la puissance des analyses de big data dépend des données à l'aide desquelles elles fonctionnent. Les données peuvent également être incomplètes, erronées ou sans valeur.

[5] La valeur réside dans les conceptions et les connaissances qui découlent des analyses de big data.

[6] La présentation de données sous une forme lisible et compréhensible, parce que la présentation des résultats d'analyses de big data peut avoir une influence énorme sur les décisions des personnes qui étudient les résultats.

[7] Données dont la signification change constamment, comme par exemple dans des langues où les mots n'ont pas toujours une même signification statique.

[8] Ceci a notamment pour conséquence que les termes "Big Data" peuvent aussi se rapporter à de petits ensembles de données. Inversement, cela signifie que tous les grands ensembles de données ne sont pas typiquement du big data.

[9] La différence entre corrélation et causalité revêt une importance cruciale. Une corrélation entre X et Y signifie qu'il existe une relation entre la variation des grandeurs X et Y (mathématiquement quantifiée par un coefficient de corrélation). Cela ne signifie toutefois pas que des modifications dans X sont aussi une cause de modifications dans Y (ou inversement). Une corrélation n'implique donc pas automatiquement une causalité. Il peut en effet aussi exister une corrélation entre X et Y parce que, par exemple, les deux dépendent d'une troisième grandeur Z. Il pourrait ainsi exister un lien entre le fait de manger des crèmes glacées et le nombre de morts par noyade, mais cela ne signifie pas pour autant que des gens se noient du fait qu'ils mangent des crèmes glacées. Plus probablement, le fait de manger des crèmes glacées a une corrélation avec la noyade en raison d'un autre élément, comme la belle saison estivale.

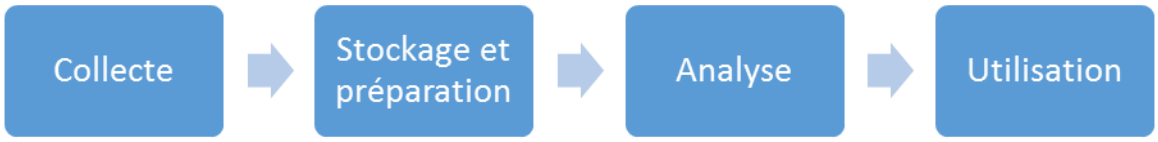
scientifique classique qu'elle est mue par des hypothèses, on parle pour le big data d'une approche mue par des données.

Enfin, on peut aussi constater que l'utilisation du big data offre de nouvelles possibilités. L'utilisation de plus grandes quantités de données peut en effet mener à de meilleures perspectives, plus détaillées, conduisant à leur tour à des processus décisionnels ou à des projections améliorés.

En résumé - et à l'instar du Wetenschappelijke Raad voor het Regeringsbeldeid (WRR) néerlandais[10] – nous pouvons déclarer qu'il nous faut considérer le big data comme une conjonction ou une convergence de facteurs (technologiques), dont celle susmentionnée, plutôt que comme une donnée bien déterminée et définissable.

B. La chaîne de valeur ("value chain") de big data

Les données en soi n'ont pas tellement de valeur. La valeur n'apparaît qu'au cours d'un processus de collecte, de préparation et de stockage, d'analyse et d'utilisation. Cela s'appelle la chaîne de valeur ou la "value chain" de big data. Nous approfondissons ci-après les différentes phases de cette chaîne de valeur.



Il est toutefois important de remarquer que les phases ne se succèdent pas toujours de manière séquentielle, mais que le processus se déroule souvent en boucles. Une analyse peut par exemple déboucher sur le choix d'autres sources de données ou sur la révision de la préparation. Il s'agit donc d'un processus plutôt répétitif que séquentiel. L'automatisation des différentes phases à l'aide d'algorithmes[11] joue à cet égard un rôle de plus en plus important par rapport à avant, bien qu'une surveillance experte et un encadrement humain demeurent nécessaires tout au long du processus.

B.1. Collecte

Suite aux énormes progrès réalisés dans le domaine ICT, on traite aujourd'hui beaucoup plus de données qu'auparavant. Grâce à la numérisation, la collecte et le stockage d'informations sont devenus à la fois plus simples et moins coûteux. Alors qu'avant, les analyses devaient être effectuées de manière ciblée et les données collectées manuellement, les données sont aujourd'hui généralement automatiquement produites de façon numérique, souvent simplement en tant que sous-produit d'opérations quotidiennes.

Ainsi par exemple, chaque transaction avec votre carte bancaire est enregistrée. Vos habitudes de navigation sur Internet sont souvent enregistrées avec précision à l'aide de cookies, de super cookies ou de la technique du "browser

[10] WRR-rapport 95 (2016). *Big Data in een vrije en veilige samenleving*, pg. 35.
[11] Voir la définition du terme algorithme dans le lexique.

fingerprinting". Nous générons en outre d'énormes quantités de données par l'utilisation des moteurs de recherche, des messageries électroniques, des médias sociaux, de la télévision numérique, de la téléphonie mobile et des applications. De plus, nous utilisons de plus en plus souvent des appareils connectés à Internet et produisant des données en temps réel - ce que l'on appelle l' "Internet des objets". Il n'y a cependant pas qu'en ligne que beaucoup de données sont collectées, une grande quantité est également recueillie hors ligne, par exemple via des enquêtes ou des cartes de fidélité (qui sont ensuite numérisées). Il est important d'observer à cet égard que les données qui sont collectées à la fois en ligne et hors ligne concernent généralement des données à caractère personnel qui, en raison de leur nature et de l'échelle de la collecte, génèrent une image souvent intrusive des personnes concernées, surtout si différentes sources sont fusionnées.

Toutes ces données sont soit collectées par des entreprises ou des administrations publiques, soit obtenues (contre un certain montant ou pas) auprès d'autres entreprises ou administrations ou de "data brokers"[12]. Les données sont en outre souvent partagées et parfois même rendues librement accessibles au public en tant qu' "open data", ce qui ne s'effectue pas toujours conformément à la législation sur la protection de la vie privée (voir le point C ci-après).

Cette diversité de sources de données résulte en un tsunami de données. Pour collecter plus de données, nous devons toutefois généralement accepter un compromis : celui de revoir un peu à la baisse nos exigences concernant l'exactitude des données. Une grande partie des données collectées sont en effet des données "brutes" pouvant en outre contenir beaucoup de bruit ; il n'est dès lors pas si simple d'en extraire des informations (données ayant du sens) ou des connaissances.

Enfin, ce qui est caractéristique du big data c'est que l'on collecte un maximum de données via toutes sortes de sources, sans savoir précisément à l'avance ce que l'on va en faire (voir le point H ci-après). Le point de départ du big data semble souvent être : au plus de données, au mieux. C'est seulement ensuite que l'on examine comment en tirer les meilleures informations. Les collectes de données peuvent donc souvent servir plusieurs finalités, qui ne deviennent en outre définies ou claires qu'au moment où ces collectes de données sont utilisées. Les données comportent en effet souvent des informations "cachées" qui n'ont de prime abord rien à voir avec la finalité initiale de la collecte, voir avec la signification des variables collectées (par exemple la déduction du risque de décès à partir de l'historique de la carte de crédit, l'orientation sexuelle à partir de connexions à Facebook, le risque de divorce à partir de l'activité sur des réseaux sociaux, le risque de comportement criminel à partir de tweets, etc.). Il s'agit d'une différence fondamentale par rapport à avant, lorsque des données étaient davantage collectées en fonction d'une seule finalité clairement définie et préalablement déterminée, bien qu'une utilisation secondaire (traitement ultérieur) était évidemment déjà possible aussi. Ceci est directement lié au fait qu'auparavant, la collecte de données représentait une entreprise complexe et coûteuse.

[12] Pour une analyse des activités des data brokers (courtiers en données), voir par exemple le rapport de la Federal Trade Commission (FTS) : *Data Brokers. A Call for Transparency and Accountability*. (mai 2014)

B.2. Stockage et préparation

La deuxième étape dans la chaîne de valeur est le stockage des données. Parallèlement aux possibilités de collecter des données, la capacité de stockage a également augmenté tandis que les coûts baissaient. Pourtant, le stockage de grandes quantités de données reste un défi, vu que la production de données augmente plus vite que la capacité de stockage.[13] La solution pour faire face à ce problème est le stockage distribué : la conservation à différents endroits de données liées entre elles sur le fond ou sur la forme. Cela revient souvent à une forme de stockage de données dans le cloud. Dans ce contexte, la sécurité des données représente un point d'attention supplémentaire.[14]

Le fait que des données soient stockées de manière distribuée ainsi que la variété des données requièrent également une autre manière d'utiliser les données. Les bases de données traditionnelles, que l'on appelle "relationnelles", ont été conçues pour un monde où les données sont peu nombreuses, structurées, prédéfinies et exactes. Cette approche est toutefois en contradiction de plus en plus flagrante avec la réalité faite de quantités sans cesse croissantes de données de type et de qualité variables, réparties dans différents disques durs et ordinateurs. La nouvelle réalité a débouché sur de nouveaux projets de bases de données où les données sont stockées de manière distribuée, généralement (semi-)déstructurée et consultées au moyen de nouveaux mécanismes, spécialement adaptés à cet effet (NoSQL⇔ SQL[15]).

Une structure logicielle populaire en la matière est Apache Hadoop, une implémentation open-source de la structure MapReduce, permettant le traitement parallèle et distribué de données à grande échelle.

Une fois que les données sont stockées, elles ne sont toutefois pas encore immédiatement prêtes pour l'analyse. Pour cela, les données doivent d'abord encore être préparées. Cela commence par la fusion de différentes sources de données pour former ce que l'on appelle un "data warehouse". Dès que ce "datawarehouse" est constitué, les données qu'il contient doivent être nettoyées et filtrées et éventuellement suivre plusieurs autres étapes de "prétraitement" (comme par exemple la normalisation ou la transformation). Ainsi, en fonction de l'application spécifique, les caractéristiques non essentielles sont par exemple supprimées, les valeurs telles que des dates et des numéros de téléphone sont normalisées, les incohérences dans les données sont éliminées autant que possible, et les doublons et les données peu fiables sont éliminés.

Dans cette phase de préparation, des techniques favorables à la vie privée sont également appliquées afin que, là où c'est nécessaire et dans la mesure du possible, on puisse éviter au maximum que des données soient reliées à des individus identifiables (voir aussi le point C ci-après). Il s'agit de techniques

---

[13] International Data Corporation,The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, april 2014, publié à l'adresse https://www.emc.com/leadership/digital-universe/2014iview/index.htm.
[14] Voir aussi à cet égard l'avis n° 10/2016 de la Commission relatif au recours au cloud computing par les responsables du traitement.
[15] SQL est l'abréviation de Structured Query Language (langage de requêtes structuré). Il s'agit d'un langage informatisé servant à interroger et à adapter des bases de données relationnelles, où les données sont stockées dans une structure en tables. Les bases de données NoSQL (Not Only SQL) offrent par contre un mécanisme permettant d'accéder à des données qui n'ont pas nécessairement été enregistrées d'une manière traditionnelle, relationnelle.

telles que la pseudonymisation[16], l'ajout de bruit, la substitution, l'agrégation ou K-anonymity, la L-diversity, la confidentialité différentielle et le hachage/la tokenisation.[17] La réserve nécessaire est toutefois de mise quant à l'efficacité de ces techniques. Toutes ces techniques ne génèrent pas des données qui ne sont plus des données à caractère personnel, c'est-à-dire des données anonymes. Des recherches ont en effet démontré qu'il était parfois relativement facile d'identifier des individus de manière unique à partir de données apparemment anonymes en utilisant des connaissances concernant des individus que la personne qui souhaite procéder à la réidentification possède parfois déjà. [18] Il est important d'étudier l'efficacité des techniques employées afin de déterminer s'il peut vraiment être question de données anonymes qui ne relèvent pas du champ d'application du Règlement.

B.3. Analyse

La troisième étape dans la chaîne de valeur est la phase d'analyse. On parle parfois aussi ici de "Knowledge Discovery in Databases" (KDD - découverte de connaissances dans les bases de données) ou de data mining[19] (exploration de données). Outre la KDD et le data mining, il existe encore beaucoup d'autres termes associés aux analyses de Big Data comme le profilage, le clustering (partitionnement de données), le text mining (fouille de textes), le machine learning (apprentissage automatique), l'analyse des réseaux (sociaux), l'analyse prédictive, le traitement automatique du langage naturel et la visualisation. Des algorithmes automatisent des décisions dans des traitements (semi-)automatisés, comme pour la fixation de priorités (moteurs de recherche, analyse des risques), le classement de données (et de personnes), l'établissement d'associations (corrélation ou causalité), le filtrage (élimination d'informations) [20] . Cependant, d'un point de vue technique et juridique, l'intention devrait être qu'un facteur humain continue à jouer un grand rôle significatif, surtout dans les cas où il peut y avoir des conséquences importantes pour des individus (voir ci-après le point N). L'homme reste, techniquement et socialement parlant, nécessaire pour concevoir, encadrer et diriger l'ensemble du processus ainsi que pour interpréter les résultats de l'analyse (ce qui représente même dans la plupart des cas un travail assez laborieux). Le data mining reste un métier. Ici, les solutions miracles ou prêtes à l'emploi n'existent généralement pas, contrairement à ce que certains fournisseurs de logiciels de produits de data mining voudraient nous faire croire.

Les analyses de big data ont pour objectif de découvrir ou de prévoir des schémas ou des caractéristiques à partir d'ensemble de données. Ces schémas n'expriment cependant que de simples corrélations, pas nécessairement des liens de causalité (voir le point D ci-après). C'est seulement en deuxième instance que ces corrélations peuvent faire l'objet d'une analyse approfondie de leur validité/reproductibilité et/ou de mécanismes de causalité sous-jacents. En

---

[16] Voir le lexique.

[17] Pour un aperçu, ainsi que des forces et des faiblesses des techniques : voir l'avis 05/2014 du G29 sur les techniques d'anonymisation (Opinion 05/2014 on Anonymisation Techniques) (WP2016).

[18] Voir par exemple de Montjoye et al (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports, 3:1376.* doi:10.1038/srep01376; de Montjoye Y. A., *Unique in the shopping mall: On the re-identifiability of credit card metadata*, Science vol 347, 30 janvier 2015; http://science.sciencemag.org/content/347/6221/536, 537.

[19] Voir le lexique.

[20] DIAKOPOULOS, N., *Accountability in Algorithmic Decision Making*, février 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/2016/03/Accountability-in-algorithmic-decision-making-Final.pdf.

outre, les corrélations ne donnent des indications que sur des liens statistiques qui sont (peut-être) réalistes au niveau de la population mais il est parfaitement possible que pour des cas individuels, ces liens généraux ne se manifestent pas.

Dans le cadre des analyses, on peut établir une distinction entre le "supervised learning" (apprentissage supervisé) et le "unsupervised learning" (apprentissage non supervisé).

Dans le cas du "supervised learning", les analyses se basent sur un modèle qui doit être éprouvé (voir aussi le schéma 1 au point G). Le modèle est éprouvé à l'aide d'un "trainingset" (ensemble d'apprentissage) de données. Cet ensemble d'apprentissage se compose des valeurs des variables collectées à propos d'un objet déterminé (par exemple des paramètres médicaux d'un patient) et du résultat ou du statut anticipé, représentés au moyen notamment d'étiquettes de classification, que l'on souhaite prédire à l'aide de ces variables (par exemple la présence ou l'absence d'un diagnostic médical). Pour concevoir un modèle, il faut autrement dit d'abord disposer d'un groupe d'objets où le résultat que l'on souhaite prédire est déjà connu. Il également important que la performance du modèle (cf. faux positifs et faux négatifs) soit vérifiée à l'aide de données de test indépendantes et représentatives. Dès que le modèle est éprouvé et testé, on peut l'utiliser pour prédire le résultat d'autres objets, dont le résultat n'est pas encore connu.

Si l'on ne dispose pas d'un ensemble d'apprentissage (donc si aucun groupe d'objets n'est disponible avec des étiquettes de classification connues), on peut alors recourir à l' "unsupervised learning" (apprentissage non supervisé). Dans ce contexte, on recherche une structure inconnue dans les données. Un exemple d'une telle technique est le clustering ou partitionnement de données, où des algorithmes recherchent des groupes, des partitionnements ou des classes dans les données.

Dans ce contexte, il en outre également utile d'établir une distinction entre les modèles "white-box" et les modèles "black-box". Ces termes connaissent différentes interprétations. De manière générale, on constate que les prévisions réalisées par des modèles black-box peuvent difficilement être expliquées de manière intuitive et/ou consistent en une relation mathématique entre input et output n'ayant aucun rapport avec les véritables lois qui régissent la relation entre input et output. Pour les modèles white-box, la base sur laquelle les prévisions s'effectuent est bel et bien interprétable et/ou basée sur les lois régissant le processus décrit par le modèle.

B.4. Utilisation

La quatrième et dernière étape dans la chaîne de valeur concerne l'utilisation des résultats de l'analyse.

Les modèles, schémas ou corrélations distillés à partir des données permettent d'avancer des hypothèses et/ou de prévoir des comportements, des caractéristiques et/ou des événements avec une précision plus ou moins grande. L'utilisation de plus grands ensembles de données fait aussi augmenter le pouvoir statistique ("power") et permet d'obtenir plus rapidement (déjà

même en cas de corrélations ou de liens subtiles ou minimaux) des résultats statistiquement significatifs (mais dans ce contexte, statistiquement significatif ne signifiera pas donc pas toujours pertinent car des liens minimaux et insignifiants sont également retenus).

Les résultats de techniques de data mining sur le big data ne sont souvent générateurs que d'hypothèses et doivent presque toujours être vérifiés ou validés (aussi bien au niveau général ou du modèle qu'au niveau individuel). En réalité, ils ne représentent donc qu'une étape intermédiaire. Les décisions automatisées doivent donc être prises avec la plus grande prudence, en particulier pour des décisions importantes ayant un impact sur la personne concernée comme en ce qui concerne la santé, l'octroi d'un crédit ou en matière d'emploi (voir le point N ci-après). Dans le cas de l'apprentissage supervisé ("supervised learning") par exemple, il faut partir du principe qu'une partie des prévisions est toujours fausse. Autrement dit, des erreurs de classification se produisent (aussi appelées faux positifs et faux négatifs), il s'agit d'objets qui sont attribués au mauvais groupe par le modèle mathématique. Un autre exemple est que dans le cadre du big data, il est possible d'analyser simultanément une quantité très importante de corrélations ou de liens. En statistique, cela s'appelle les tests multiples ("multiple testing") et cela a pour conséquence que l'on voit parfois des liens qui n'existent pas, par pur hasard. Le surapprentissage du modèle (overfitting) est un autre exemple où l'on modélise des liens qui n'existent pas, ce qui dégrade la performance du modèle sur des données de test indépendantes.

En dépit des observations critiques ci-dessus, des décisions et des prédictions plus précises obtenues grâce au data mining peuvent à leur tour potentiellement mener notamment à la réduction d'erreurs humaines, un accroissement de l'efficacité opérationnelle, une réduction des coûts et des risques, de nouveaux produits, une plus-value sociale (par exemple dans les soins de santé) et à une optimisation des offres. Ceci explique le grand intérêt manifesté pour les applications et les possibilités du big data tant par les entreprises que par les pouvoirs publics.

Exemples :

- Google Flu Trends : Google prévoit les épidémies de grippe sur la base de milliards de recherches.
- Amazon fait des suggestions pour des livres sur la base d'algorithmes de big data en utilisant le propre profil d'achat et celui d'autres personnes.
- Les autorités publiques peuvent utiliser le big data pour détecter diverses formes de fraude[21].
- Antiterrorisme (cf. le programme PRISM de la NSA) et police prédictive (predictive policing)[22] (le projet iPolice de la Police fédérale[23]).

---

[21] Voir l'Accord de gouvernement du 9 octobre 2014, http://www.premier.be/sites/default/files/articles/Accord_de_Gouvernement_-_Regeerakkoord.pdf ; TOMMELEIN, B., Exposé d'orientation politique du secrétaire d'État à la Lutte contre la Fraude sociale, à la Protection de la Vie privée et à la Mer du Nord du 13 novembre 2014, Chambre, DOC 54, 0020/004, publié sur http://www.dekamer.be/FLWB/PDF/54/0020/54K0020004.pdf.

[22] Voir par exemple les références aux applications par le Los Angeles Police Department à la page 20 de *Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights*, mai 2016, publié sur https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

- Publicités ciblées ("Online Behavioural Advertising") : l'analyse des habitudes de navigation, de cliquage, de visionnage [24] génère beaucoup d'argent car les publicitaires sont prêts à payer plus pour montrer des publicités ciblées.
- Sciences médicales : prévisions d'effets thérapeutiques et d'effets secondaires de médicaments et d'autres interventions thérapeutiques, soutien des médecins dans le choix du traitement (clinical decision support), diagnostic ciblé et plus précis, médecine personnalisée (personalised medicine) ...
- Bio-informatique : de nouvelles techniques en biologie moléculaire/génétique permettent de générer de très grandes quantités de données. Il est par exemple possible actuellement de déterminer le génome complet (next generation sequencing) d'un individu pour un prix abordable. Il est par exemple également possible depuis un certain temps de déterminer le transcriptome (concentration de toutes les molécules d'ARN dans une cellule) de cellules tumorales, notamment. Cette forme de big data combinée à des techniques de data mining permet d'établir des schémas thérapeutiques personnalisés où le traitement correspond beaucoup mieux au profil génétique du patient (ou de la tumeur).

---

[23] MEEUS, R., Longread: *Hoe iPolice de natie veiliger maakt*, 24 juin 2015, http://datanews.knack.be/ict/nieuws/longread-hoe-ipolice-de-natie-veiliger-maakt/article-longread-720899.html ; Ponciau, L, *Les détails du projet iPolice dévoilés*, Le Soir, 17 septembre 2016.

[24] *Proximus TV gaat reclame aanpassen aan uw kijkgedrag*, Express Business, 1er août 2016 ; *Proximus ouvre la voie à la publicité personnalisée*, le Soir, 2 août 2016, *Telenet begint met tv-reclame op maat*, De Tijd, 13 septembre 2016.

Partie 2

Principes de protection des données et recommandations concernant les analyses de big data

A. Remarques générales - méthodologie

La législation sur la protection des données à caractère personnel s'applique dès qu'il est question d'un traitement de données à caractère personnel[25]. Pour qu'il soit question d'un traitement de données à caractère personnel, il suffit que des personnes puissent être isolées dans des ensembles de données (voir la référence au "single out" au point C ci-après), entraînant un risque raisonnable qu'un aspect de leur identité puisse être retrouvé (il peut aussi s'agir d'une identité numérique, comme par exemple une adresse e-mail, permettant d'approcher la personne concernée)[26], sans qu'il soit nécessaire de connaître absolument leur nom ou leur identité civile.

Dans le cadre de l'obligation de définir explicitement la finalité du traitement (article 5.1.b) du RGPD) et de l'approche basée sur les risques ("**risk based approach**")[27] du RGPD (impliquant l'obligation de vérifier l'impact sur les droits et libertés des personnes concernées), il est important de toujours vérifier quel effet le recours à des analyses de big data peut avoir dans la pratique à l'égard des personnes concernées.

Il y a une différence entre la nature des analyses de big data et leur utilisation. En ce qui concerne l'utilisation, on peut établir une distinction entre l'utilisation prédictive[28], descriptive[29] ou prescriptive[30] d'analyses de big data[31]. De telles utilisations peuvent avoir une influence sur les modalités de traitement et sur les droits et libertés des personnes concernées (voir ci-après au point E.2.).

Les principes de nombreux projets de big data sont en contradiction avec les principes de respect de la vie privée et de protection des données[32], la modularité[33] et l'approche basée sur les risques en vertu du RGPD (voir ci-

---

[25] Voir les définitions à l'article 1, §§ 1 et 2 de la loi du 8 décembre 1992 et à l'article 4, 1) du RGPD.

[26] Voir le considérant 26 du RGPD dont la version anglaise renvoie au "singling out", et l'Avis du 08/2012 WP 199 du Groupe 29, publié sur http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp199_en.pdf. D'après l'arrêt de la Cour de Justice de l'Union européenne dans l'affaire C-582/14 du 19 octobre 2016 Patrick Beyer/Bundesrepublik Deutschland, "*l'adresse de protocole internet dynamique d'un visiteur constitue, pour l'exploitant du site, une donnée à caractère personnel, lorsque cet exploitant dispose de moyens légaux lui permettant de faire identifier le visiteur concerné grâce aux informations supplémentaires dont dispose le fournisseur d'accès à Internet du visiteur.*"

[27] Voir WP 218, Déclaration du Groupe 29 du 30 mai 2014 sur le rôle d'une approche basée sur les risques dans des cadres juridiques de protection des données, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

[28] Comme il apparaît ci-après, ce sont surtout les analyses prédictives et prescriptives qui présentent certaines caractéristiques sensibles pour la vie privée et un risque de traitement déloyal dû à la discrimination.

[29] On entend ci-après par "modèles descriptifs" : l'étude de corrélations ou de schémas dans des traces digitales. Dans le cas des modèles prédictifs, on tente - en utilisant ces corrélations - de prévoir une (des) variable(s) d'exportation (output) (par exemple le risque de fraude) à l'aide de variables d'importation (input).

[30] Utilisation de modèles mathématiques pour orienter le comportement (d'achat ou conforme aux normes) (par exemple décourager des achats hors ligne en proposant une alternative en ligne plus avantageuse qui laisse plus de traces, comme pour les programmes de fidélisation de la clientèle des grandes surfaces, ou des mesures stratégiques empêchant l'achat anonyme de cartes prépayées).

[31] Voir page 21 de MOEREL, Lokke et PRINS, Corien, *Privacy for the homo digitalis*. Proposal for a new regulatory framework for data protection in the light of Big Data and the Internet of Things, 25 mai 2016, disponible sur http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784123, en de definities hierna en VAN DER SLOOT, B. et VAN SCHENDEL, S. *Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data*, p. 40, publié à l'adresse http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf.

[32] International Working Group on Data Protection in Telecommunications, Working paper on Big Data and Privacy. Privacy principles under pressure in the age of Big Data analytics, 5-6 mai 2014, publié sur http://www.datenschutz-berlin.de/attachments/1052/WP_Big_Data_final_clean_675.48.12.pdf.

[33] Cette "modularité" signifie que pour les traitements présentant un risque élevé, il conviendra de prendre davantage de mesures pour respecter les exigences de protection des données que pour les traitements présentant un faible risque. L'idée illustre un glissement de l'accent qui était mis auparavant sur la collecte de données à caractère personnel vers

après).

Il existe un besoin manifeste de plus de **sécurité (juridique)** concernant la question de savoir si et comment les règles de respect de la vie privée et de protection des données présentes et à venir[34] doivent être appliquées.

La Commission considère comme un défi de conférer à la nouvelle approche en vertu du RGPD et aux principes classiques de respect de la vie privée et de protection des données à caractère personnel un effet d'utilité dans le contexte d'analyses de big data. À cet égard, l'intention ne peut pas être de freiner inutilement l'utilisation de ces nouvelles technologies de l'information, vu leur utilité souvent avérée pour la société.

Comme mentionné ci-avant dans la partie I, la **réglementation doit être générale, solide et suffisamment neutre technologiquement** (ce qu'elle est d'ailleurs pour une grande partie) pour que lors de tout traitement, la vie privée des personnes concernées soit suffisamment garantie et qu'un **bon équilibre** soit trouvé afin de ne pas trop entraver les possibilités légitimes, du fait que les conditions permettant un usage légitime sont offertes.

Pour relever ce défi, la Commission formule ci-après **des recommandations concrètes** qui peuvent être utilisées afin de vérifier et de favoriser le degré de respect des règles de protection de la vie privée et des données lors d'analyses de big data. Pour répondre aux préoccupations de certains acteurs concernés d'atteindre une situation équitable par rapport à d'autres pays européens, une consultation publique a eu lieu en **avril 2017** au sujet de ces recommandations. On a également tenté de les rédiger **en se basant sur l'application du RGPD.** Ces recommandations s'appliquent sans préjudice des points de vue du Comité européen de la protection des données.

Lors de l'application des règles de protection de la vie privée et des données à caractère personnel, il est important de connaître les différentes phases[35] qui peuvent être distinguées dans le cadre de projets de big data. La différence est en effet essentielle selon que les principes de respect de la vie privée et de protection des données sont appliqués sur toutes les phases susmentionnées plutôt que sur une seule partie d'un projet de big data.

B. RGPD

La Commission attire l'attention sur le fait qu'une nouvelle réglementation européenne relative à la protection des données à caractère personnel a été promulguée récemment : le Règlement général relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données et la Directive Police et Justice[36]. Ces

---

un positionnement sur l'utilisation des données à caractère personnel, comme dans les discussions relatives au big data. Voir WP 218, Déclaration du Groupe 29 du 30 mai 2014 sur le rôle d'une approche basée sur les risques dans des cadres juridiques de protection des données, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

[34] Voir la référence au RGPD au point B ci-après.

[35] Voir le renvoi dans la partie I ci-avant aux notions de "collecte", "stockage et préparation", "analyse" et "utilisation".

[36] Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 *relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de*

textes ont été publiés au journal officiel de l'Union européenne le 4 mai 2016[37].

Le Règlement, couramment appelé RGPD (pour Règlement général sur la protection des données), est entré en vigueur vingt jours après sa publication, soit le 24 mai 2016, et est automatiquement applicable deux ans plus tard, soit le 25 mai 2018. La Directive Police et Justice doit être transposée dans la législation nationale au plus tard le 6 mai 2018.
Pour le RGPD, cela signifie que depuis le 24 mai 2016, pendant le délai d'exécution de deux ans, les États membres ont d'une part une obligation positive de prendre toutes les dispositions d'exécution nécessaires, et d'autre part aussi une obligation négative, appelée "devoir d'abstention". Cette dernière obligation implique l'interdiction de promulguer une législation nationale qui compromettrait gravement le résultat visé par le Règlement. Des principes similaires s'appliquent également pour la Directive.

Il est dès lors recommandé d'anticiper éventuellement dès à présent ces textes. Dans le présent rapport, la Commission a d'ores et déjà veillé, dans la mesure du possible et sous réserve d'éventuels points de vue complémentaires ultérieurs, au respect de l'obligation négative précitée.

Bien que le RGPD ne contienne pas de définition de "big data" en tant que tel, ce Règlement introduit une série de nouveaux éléments qui paraissent pertinents dans le contexte des analyses de big data. Sans vouloir être exhaustive, la Commission renvoie à la définition du profilage dans le RGPD[38] qui, à l'instar d'une recommandation précédente du Comité des Ministres du Conseil de l'Europe[39], met l'accent sur le caractère analytique et prédictif de cette technique. Le RGDP adopte une approche basée sur les risques ("risk based approach") (couplée à l'analyse d'impact relative à la protection des données[40] (voir ci-après les points E et S)). Le RGPD contient aussi de nouvelles obligations et de nouveaux principes tels que la responsabilité ("accountability"[41]) (voir ci-après au point S), les obligations de notification[42] et de consultation pour les risques résiduels[43], la constitution d'une documentation

---

sanctions pénales, et à la libre circulation de ces données et abrogeant la décision-cadre 2008/977/JAI du Conseil (JO L 119, 4 mai 2016, p. 89–131).

[37] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 *relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données)*.
Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 *relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données et abrogeant la décision-cadre 2008/977/JAI du Conseil*.
http://eur-lex.europa.eu/legal-content/NL/TXT/?uri=OJ:L:2016:119:TOC
http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ%3AL%3A2016%3A119%3ATOC

[38] L'article 4, 4) du RGPD contient la définition suivante du profilage : "*toute forme de traitement automatisé de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique*

[39] *Voir aussi la définition à* l'article 1. e. de la Recommandation CM/Rec(2010)13 du 23 novembre 2010 du Comité des Ministres aux États membres *sur la protection des personnes à l'égard du traitement automatisé des données à caractère personnel dans le cadre du profilage*, publiée à l'adresse :
https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016805cdd0e.

[40] Article 35 du RGPD. L'analyse d'impact relative à la protection des données sera (notamment) nécessaire si un projet de big data implique l'évaluation systématique et approfondie des aspects personnels concernant des personnes physiques, sur base de laquelle sont prises des décisions produisant des effets juridiques à l'égard d'une personne physique ou l'affectant de manière significative de façon similaire.

[41] Article 5.2 du RGPD.

[42] Articles 33 et 34 du RGPD.

[43] Article 36.1 du RGPD.

interne[44], les principes de protection des données dès la conception ("privacy by design")[45] et de protection des données par défaut ("privacy by default")[46] ainsi que le rôle des codes de conduite et de la certification[47] (les codes de conduite et la certification peuvent assurément jouer un rôle important pour la protection des données à caractère personnel dans le contexte de big data – ces concepts feront l'objet de développements ultérieurs dans des avis complémentaires des contrôleurs aux niveaux européen ou national).

On peut en outre encore mentionner : la transparence (en tant que principe général)[48], les conditions applicables au consentement[49], le règlement des données sensibles[50] et le droit à l'information[51] (par exemple existence d'une prise de décision automatisée).
Parmi les éléments nouveaux, on compte également le fait que le RGPD considère l'évaluation d'aspects personnels, auxquels sont couplées des conséquences juridiques pour la personne concernée ou qui touchent considérablement cette dernière[52], comme un **risque** pour les droits et libertés des personnes physiques concernées. Il en résulte que le responsable du traitement doit effectuer de manière objective[53] une **surveillance continue des risques**[54], la méthode d'évaluation des risques **devant être réévaluée périodiquement**. Il sera (notamment)[55] question d'un **risque élevé** s'il y a une évaluation large et systématique d'aspects personnels[56].

Si par exemple un modèle d'Ipolice analyse les réactions de citoyens sur les médias sociaux en vue d'exécuter des missions de police administrative ou judiciaire, l'analyse d'impact devra prendre en considération le risque et l'influence au niveau de la liberté d'expression des personnes concernées sur les médias sociaux ainsi que le traitement équitable (interdiction de discrimination) par la police, compte tenu des mesures de contrôle déjà existantes sur les services de police. Selon les applications, il faudra également mettre en balance l'impact sur le droit à la propriété[57] et la liberté individuelle (d'application dans le

---

[44] Article 30 du RGPD (registre des activités de traitement), article 33.5 du RGPD (pour toutes les violations relatives aux données à caractère personnel).
[45] Article 25 du RGPD.
[46] Article 25 du RGPD.
[47] Article 40 du RGPD.
[48] Considérants 38, 58 et 100, article 51. a) et 88 2. du RGPD.
[49] Voir notamment les articles 4 11), 7 et 8 du RGPD ainsi que les considérants n° 32, 33, 38, 40, 42, 43 et 50.
[50] Voir les articles 9 et 10 du RGPD ainsi que les considérants n° 10, 34, 35, 54-54, 71, 80, 91 en 97.
[51] Voir les articles 12 à 15 inclus, 19, 21.4 du RGPD ainsi que les considérants n° 58, 60-63, 71, 156.
[52] Considérant 71 du RGPD.
[53] Le considérant 76 du RGPD confirme le caractère objectif de cette évaluation du risque à l'égard du traitement et des ses conséquences pour les droits et libertés des personnes : "*Il convient de déterminer la probabilité et la gravité du risque pour les droits et libertés de la personne concernée en fonction de la nature, de la portée, du contexte et des finalités du traitement. Le risque devrait faire l'objet d'une évaluation objective permettant de déterminer si les opérations de traitement des données comportent un risque ou un risque élevé.*"
[54] Voir la classe de risques mentionné au considérant 75 du RGPD et notamment les articles 32.1, 33, 34 et 35 du RGPD.
[55] Pour une liste des risques qui doivent être qualifiés d'élevés : voir Groupe 29, WP 248, - Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 (du 4 avril 2017), publié à l'adresse http://ec.europa.eu/newsroom/document.cfm?doc_id=44137 ; et l'annexe 2 du Projet de recommandation de la CPVP *concernant l'analyse d'impact relative à la protection des données*, publié à l'adresse https://www.privacycommission.be/sites/privacycommission/files/documents/CO-AR-2016-004_FR.pdf.
[56] Considérant 91 du RGPD.
[57] Voir le document de travail n° 112 de la Banque nationale de 2011 qui démontrait un lien entre les retards de paiement en matière de téléphonie mobile et les retards de paiement en matière de crédit, https://www.nbb.be/en/articles/working-paper-ndeg-212. Si la Centrale des crédits aux particuliers est réformée pour travailler sur la base de telles corrélations, cela peut avoir un impact sur la possibilité d'accéder à la propriété.

cadre de la police prédictive)[58]. Le fait est d'autre part que pour l'application de l'approche basée sur les risques en vertu du RGPD[59], il existe un important besoin d'élaborer une méthode satisfaisante pour l'application de l'évaluation des risques à des analyses de big data.

L'analyse des risques appliquée devra recourir à une méthode comportant plusieurs éléments de base que la Commission expliquera plus en détail dans le cadre d'une recommandation distincte. La notion de "risque" peut être interprétée de plusieurs façons. Dans la littérature, on décrit généralement la notion de "risque" comme la possibilité qu'une menace déterminée se présente, avec pour conséquence un impact déterminé ("gravité").[60]

Dans le cas d'une évaluation continue des risques, il conviendra aussi de prévoir régulièrement un moment d'évaluation où le modèle d'évaluation des risques sera réévalué. Lors de l'analyse des risques liés à des analyses de big data (voir ci-avant), **il est préférable de distinguer et de considérer toutes les différentes phases (voir ci-avant)** [61]. Se limiter à expliquer un seul traitement (comme la collecte, l'organisation, ... de données[62]) et/ou le risque lié à un traitement (par exemple le codage de données) peut en effet donner une image incomplète des risques lors de l'application à des personnes physiques.

La Commission constate qu'il existe dans les flux professionnels ("B2B") une promotion et une vente croissantes de toutes sortes de services et de solutions liés au big data. Cette promotion s'accompagne souvent de la promesse de perspectives peut-être irréalistes[63] aux responsables, par exemple en termes d'économie de dépenses lors de l'approche de la lutte contre la fraude. On oublie à cet égard que le potentiel du big data ne peut être réalisé que s'il est au moins satisfait à une série de conditions connexes[64], dont la disponibilité et la qualité de données contenant les informations souhaitées (voir ci-après), la présence et l'application correcte de l'expertise au sein de l'organisation du responsable et l'existence d'un bon cadre réglementaire permettant à la fois des analyses de big data efficaces et la préservation des droits et libertés des personnes.

✓ Recommandation 1

Un responsable du traitement qui souhaite réagir à une offre de tiers de tester ou d'appliquer la technologie du big data peut toujours demander les informations ou la documentation décrivant l'impact de cette technologie sur la

[58]Voir ci-avant et le lexique. Un exemple est l'application au niveau du Los Angeles Police Department, décrite à la page 20 de l'Executive Office of the President : Big Data : *A Report on Algorithmic Systems, Opportunity, and Civil Rights*, mai 2016, publié sur https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
[59]Voir l'annonce de l'ENISA de janvier 2016 concernant (notamment) l'aspect de l'analyse des risques du RGPD sur ce point https://www.enisa.europa.eu/publications/enisa-position-papers-and-opinions/enisa2019s-position-on-the-general-data-protection-regulation-gdpr/.
[60] Voir par exemple I. Naumann (ed.), "Privacy and Security Risks when Authenticating on the Internet with European eID Cards", ENISA, 26 novembre 2009.
[61] Voir la partie I ci-avant : collecte, stockage et préparation, analyse et utilisation
[62] Au sens de l'article 1, § 2 de la LVP.
[63] Voir dans la partie I le renvoi au big data en tant que "métier".
[64] Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, p 84 point 4.4.

vie privée des personnes physiques ("analyse d'impact relative à la protection des données" du produit ou du service)[65] et/ou un label pertinent en matière de protection des données[66]. Ces informations ou cette documentation ne peuvent pas être une analyse statistique ni une attestation confirmant que l'utilisation de la technologie est conforme au RGPD.

Vu les principes de protection des données dès la conception ("privacy by design") et de protection des données par défaut ("privacy by default") ainsi que l'obligation de responsabilité ("accountability"), le RGPD comporte une obligation pour les responsables du traitement de pouvoir démontrer que la technologie big data qu'ils utilisent est conforme au RGPD.

✓ Recommandation 2

Toute promotion de la mise en œuvre d'une technologie big data pour des applications spécifiques[67] devrait aussi informer clairement (par exemple via un label pertinent en matière de protection des données ou dans la fiche produit)[68] les clients professionnels ("B2B") potentiels (organisations qui veulent utiliser cette technologie) pour savoir si et comment le recours spécifique à cette technologie respecte la réglementation européenne en matière de vie privée et de protection des données.
Il ne semble pas évident à première vue d'évaluer et de prouver [69] **la conformité ou non avec le RGPD ("GDPR compliance") et/ou la responsabilité dans le cadre du big data**. La présente recommandation tente en tout cas de donner une ébauche de solution. Le RGPD contient bien entendu aussi des éléments qui peuvent être pertinents à cet égard, comme le fait de tenir compte ou non du risque de certaines applications de big data pour les droits et libertés des personnes physiques[70], la garantie de transparence ("B2C") (à la personne concernée)[71], l'utilisation ou non de labels pertinents en matière de protection des données ou la conclusion de codes de conduite[72].

✓ Recommandation 3

Via l'analyse d'impact relative à la protection des données, les responsables peuvent indiquer dans quelle mesure il a été tenu compte du RGPD et

---

[65] Bien que pour éviter des exercices pro forma ("check the box"), on ne puisse pas imposer une forme ni un contenu fixes pour ces analyses, on peut toutefois se référer à plusieurs caractéristiques minimales auxquelles cette analyse d'impact doit répondre. Voir l'annexe 1 du projet de recommandation d'initiative du 20 décembre 2016 *concernant l'analyse d'impact relative à la protection des données*, publié à l'adresse https://www.privacycommission.be/sites/privacycommission/files/documents/CO-AR-2016-004_FR.pdf. Voir également les exemples et critères dans l'annexe du Groupe 29, WP 248, - Guideliness on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 du 4 avril 2017, publiée à l'adresse http://ec.europa.eu/newsroom/document.cfm?doc_id=44137.

[66] Considérant 100 et articles 28.5 et 42 du RGPD.

[67] Certaines technologies big data sont génériques et peuvent être utilisées pour différentes applications et de ce fait, il est difficile d'évaluer la conformité avec le RGPD.

[68] Une déclaration de confidentialité classique dans les conditions générales de produit ne suffira pas ici.

[69] Considérant 90 et article 24.1 du RGPD.

[70] Considérants 74, 84 et 150 du RGPD.

[71] Considérants 39, 58 et 78 et article 5.1 a) du RGPD.

[72] Considérants 98, 148 et articles 24.3 et 28.5 du RGPD.

éventuellement des présentes recommandations de la Commission.

✓ **Recommandation 4**

En cas de recours à des analyses de big data qui constituent un risque élevé[73] pour les droits et libertés des personnes concernées, il peut être recommandé d'impliquer les personnes concernées et divers acteurs concernés[74] de la société civile par le biais par exemple de mécanismes de feed-back comme des enquêtes et le cas échéant, de conclure un code de conduite afin d'encourager l'utilisation responsable de ces traitements.

C. Applicabilité éventuelle de la législation relative à la protection des données lors de l'application de techniques de pseudonymisation ou d'anonymisation, en combinaison ou non avec l'agrégation de données

On pense souvent à tort que l'utilisation de données dans un environnement de test ou dans un projet pilote ne constitue pas un traitement de données à caractère personnel relevant de la législation relative à la protection des données.

Des responsables pensent parfois trop vite que du fait que les données sont codées (pseudonymisation [75]) ou anonymisées, le droit à la protection des données ne s'y applique pas. On oublie en outre souvent que la conversion de données à caractère personnel en données (soi-disant) anonymes implique aussi un traitement de données à caractère personnel. Lors de ce que l'on appelle l' "anonymisation" et/ou l'agrégation de données, appliquée(s) dans une phase déterminée du projet, il n'existe pas toujours non plus de garanties suffisantes que la possibilité d'une réidentification soit totalement exclue. Autrement dit, on n'est pas certain qu'aucune donnée ne sera traitée concernant des personnes physiques identifiables. Parfois, les arguments dont on dispose sont donc insuffisants pour exclure complètement et définitivement l'applicabilité de la loi sur la protection des données.

On retrouve parfois le raisonnement selon lequel le fait que les données soient codées ou anonymisées doit être considéré comme une garantie suffisante pour la protection de la vie privée. Ce principe n'est pas toujours correct.

Des chercheurs ont démontré à plusieurs reprises au cours des dernières années que même des données soi-disant anonymisées ou codées[76] pouvaient être reliées sans beaucoup de difficulté à des individus (il s'agit du fait d'individualiser des personnes dans des ensembles de données ou "**to single out**" - voir aussi le lexique)[77], avec les risques de réidentification que cela

[73] Voir la note de bas de page 56 ci-avant.
[74] Industrie, consommateurs, syndicats et société civile, autorités compétentes et contrôleurs.
[75] Voir la définition à l'article 4 5) du RGPD.
[76] Sans que l'on ait dans ce cas accès à la traduction du code en un identifiant significatif.
[77] Voir la recherche citée ci-avant de De Montjoye Y.-A., Hidalgo C. A., Verleysen M. et Blondel V. D.,"Unique in the Crowd: The privacy bounds of human mobility", Nature Scientific Report 3, 25 Mars 2013, http://www.nature.com/articles/srep01376, citée à la p. 67 du Wetenschappelijke Raad voor het Regeringsbeleid, Big

implique. Certains projets - comme dans le cadre de l'analyse de données à caractère personnel financières lors d'achats[78] ou dans le cadre du traçage de télécommunications[79] ou d'activités en ligne à l'aide de cookies - sont riches en données permettant potentiellement une réidentification, bien qu'elles ne comportent pas de noms, de numéros de compte ou de numéros d'identification évidents. Les projets pour lesquels on utilise des ensembles de données riches en variables (telles que le lieu, le prix (la catégorie de prix) et le moment de l'achat qui figurent dans les données de transaction) présentant une grande probabilité qu'une combinaison de ces variables détermine un individu de manière univoque comportent un grand risque d'individualisation ("singling out") et donc par conséquent, de réidentification (il va de soi que le risque concret de réidentification dépend également des données concernant les personnes concernées auxquelles une partie - qui pourrait/souhaiterait procéder à une réidentification - peut supposément déjà (pouvoir) accéder). Tel est par exemple le cas pour les données générées par certaines applications mobiles, des ensembles de données financières, l'historique de navigation ou les données de compteurs intelligents. Dans de tels cas, l'unicité (voir lexique) des données est tout simplement trop grande. Ce n'est dès lors pas un hasard si plusieurs de ces ensembles de données bénéficient d'une protection (supplémentaire) dans le cadre du secret des communications.[80]

Dans tous ces cas, il est donc question d'un traitement de données à caractère personnel et la législation sur la protection des données est clairement applicable.

Afin de réduire ou même d'exclure le risque d'individualisation (singling out) ou de réidentification, on procède parfois à l'agrégation de données[81]. En cas d'agrégation, on peut souvent argumenter à juste titre que - vu que la réidentification n'est plus possible - la législation sur la protection des données ne s'applique pas (ne s'applique plus) (à une partie du projet), par exemple lors de la communication de statistiques à des tiers. Toutefois, lorsque le résultat de l'agrégation est ensuite utilisé à l'égard de personnes[82] (par exemple lors de l'utilisation de corrélations – les corrélations peuvent être considérées comme une forme d'informations résumées ou agrégées), il y a (à nouveau) un (autre) traitement de données à caractère personnel. La législation relative à la protection des données s'appliquera à nouveau dans ce cas, même s'il est (partiellement) question de l'utilisation de données anonymes ou agrégées.

Outre l'agrégation, il existe encore d'autres techniques pour anonymiser des données (telles que la randomisation (ajout de bruit, permutation, confidentialité différentielle) et la généralisation (dont relèvent les techniques

---

Data in een vrije en veilige samenleving, University Press Amsterdam, publié sur http://www.wrr.nl/fileadmin/nl/publicaties/PDF-Rapporten/rapport_95_Big_Data_in_een_vrije_en_veilige_samenleving.pdf.

[78] De Montjoye Y. A., Unique in the shopping mall: On the re-identifiability of credit card metadata, Science 347, 30 janvier 2015 ; http://science.sciencemag.org/content/347/6221/536, 537.

[79] Voir De Montjoye Y.-A., Hidalgo C. A., Verleysen M. and Blondel V. D.,"Unique in the Crowd: The privacy bounds of human mobility", Nature Scientific Report 3, 25 mars 2013, http://www.nature.com/articles/srep01376.

[80] Voir l'article 5 de la Directive 2002/58/CE du 12 juillet 2002 qui protège la confidentialité des télécommunications ainsi que l'article 124 de la loi du 13 juin 2005 *relative aux communications électroniques*, M.B. du 20 juin 2005.

[81] Voir le lexique.

[82] La détermination d'une caractéristique (par exemple le calcul du risque de fraude à l'aide d'un résultat de synthèse découlant d'analyses de big data) et son application ensuite à une personne constitue déjà assez directement un traitement automatisé de données à caractère personnel.

d'agrégation ou de k-anonymat, mais aussi les technique de l-diversité et de t-proximité))[83]. Nous ne les approfondirons pas dans le présent rapport.

L'analyse de big data est déjà appliquée en Belgique avec l'agrégation de données (donc dépouillées de toute donnée client et de l'identifiant technique), comme par exemple l'analyse de données de localisation par les opérateurs de télécommunications belges Proximus pour le tourisme[84] et Orange pour de grands événements[85]. Le fait de travailler avec des données de localisation groupées d'au moins 30 utilisateurs finaux mobiles a été considéré par la Commission comme acceptable dans le cadre de tels projets[86].

Le fait qu'il soit question ou non d'agrégation dans un projet de big data peut constituer un critère de contrôle pour à la fois évaluer l'applicabilité de la législation sur la protection des données et analyser la proportionnalité des traitements. La question de savoir si l'agrégation peut être considérée comme une garantie réelle pour répondre à l'exigence d'une ingérence proportionnelle dans la vie privée devra être examinée au cas par cas.

✓ Recommandation 5

L'agrégation, la pseudonymisation ou la (soi-disant) anonymisation de données ne suffisent pas toujours à protéger efficacement les données à caractère personnel ou à empêcher la réidentification. Il convient toujours d'évaluer cela au cas par cas. Pour pouvoir démontrer par exemple que l'agrégation de données dans une phase d'un projet de big data constitue aussi une garantie suffisante, il est préférable de coupler cette agrégation à **une analyse quantitative qui calculera le risque d'individualisation ou de réidentification** sur la base des ensembles de données (agrégées) (par exemple **"Small Cells Risk Analysis"** (analyse de risques de petits groupes de données) ou "SCRA" pour des données agrégées)**.** Vu le caractère parfois dynamique du risque de réidentification, il vaudra donc mieux dans de nombreux cas répéter cette analyse dans le temps (**garantie périodique**).

L'Analyse de Risques Small Cells (SCRA) est une méthode où le risque d'identification indirecte dans des données agrégées (dans ce rapport, il s'agit donc de données rapportant uniquement des statistiques globales pour des groupes d'individus) est évalué sur la base de la combinaison de variables "quasi ID"[87].

---

[83] Voir l'avis 05/2014 WP 216 du 10 avril 2014 sur les techniques d'anonymisation, publié sur http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_fr.pdf.

[84] http://www.proximus.be/fr/id_b_cl_tourism/grandes-entreprises-et-secteur-public/decouvrir/blog/one-magazine/temoignages/le-tourisme.html.

[85] https://business.orange.be/fr/decouvrir/nouvelles/orange-analyse-le-public-des-%C3%A9v%C3%A9nements.

[86] La Commission vie privée réagit : "Tout couplage avec quelle que donnée de client que ce soit n'est en aucun cas acceptable à nos yeux" [traduction libre réalisée par le Secrétariat de la Commission en l'absence de traduction officielle], 25 novembre 2016, Datanews, http://datanews.knack.be/ict/nieuws/privacycommissie-reageert-iedere-koppeling-met-enig-klantgegeven-is-voor-ons-onder-geen-beding-aanvaardbaar/article-opinion-781325.html.

[87] Les variables "quasi ID" sont des caractéristiques qui sont utilisées pour délimiter les groupes dans des données agrégées mais qui, en combinaison, pourraient aussi être utilisées en vue d'identifier un individu. Des exemples sont le pays du domicile, le code postal, le sexe, l'âge ou la date de naissance. Voir glossaire.
Le risque d'identification dépend du nombre et de la nature de ce type de variables dans les données et des connaissances a priori de celui qui tente de procéder à l'identification.

La SCRA est une analyse théorique (qui s'effectue sans que l'on doive déjà disposer de l'ensemble de données devant être agrégées) du nombre de combinaisons uniques des valeurs rapportées de ces variables quasi ID par rapport au nombre de points dans les données au niveau individuel.

Si le risque d'identification indirecte (en raison de la présence potentielle de small cells, c'est-à-dire de groupes au sein des données agrégées présentant un nombre de points trop faible au niveau individuel) est trop élevé, il peut être conseillé d'appliquer un certain nombre de restrictions (par exemple supprimer une ou plusieurs variables quasi ID, agréger une variable quasi ID telle que l'âge ou la catégorie d'âge, ...). Lors de l'analyse des risques, on examinera également si les statistiques globales - qui sont rapportées pour les éventuelles small cells - comportent des informations supplémentaires et/ou sensibles sur les individus figurant dans les small cells.

Une telle analyse des risques est fréquente pour les fournitures de données dans le secteur de la sécurité sociale (par exemple par l'Agence Intermutualiste (AIM[88]) et en ce qui concerne le projet Thales[89]).

Si la présence de small cells doit être évaluée au moment où l'ensemble de données qui doivent être agrégées est déjà disponible, il n'est évidemment plus nécessaire d'effectuer une analyse des risques mais il faut simplement vérifier la présence parmi les données de groupes (small cells) dont le nombre de points au niveau individuel est inférieur à un certain seuil. Si tel est le cas, on peut - et à nouveau uniquement si les statistiques globales qui sont rapportées comportent des informations supplémentaires et/ou sensibles sur les individus dans les small cells - masquer les informations relatives à ces groupes afin de garantir l'anonymat au maximum.

Cette analyse des risques des small cells doit donc évaluer l'unicité/la "granularité" des ensembles de données ou la possibilité d'individualisation ("singling out") ou de réidentification. Ce n'est que si une telle analyse démontre que la possibilité d'individualisation est suffisamment exclue que les données sont suffisamment agrégées ou anonymisées. Cette analyse des risques des small cells n'a en outre de sens que si l'on peut supposer que celui qui souhaite/peut procéder à la réidentification sait précisément qui appartient au groupe d'individus utilisé pour constituer l'ensemble de données.

✓ Recommandation 6

L'agrégation ou l'anonymisation de données dans une phase d'un projet de big data n'offrent des garanties suffisantes pour la protection des données à

[88] Voir par exemple l'analyse Small Cells de la base de données complète de l'Enquête de santé 2013. Analyse exécutée par l'Agence Intermutualiste, juillet 2015. Voir la recommandation n° 11/03 du 19 juillet 2011 du Comité sectoriel de la Sécurité Sociale et de la Santé, Section Santé, relative à une note du centre fédéral d'expertise des soins de santé relative à l'analyse Small Cells de données à caractère personnel codées provenant de l'Agence intermutualiste https://www.privacycommission.be/sites/privacycommission/files/documents/recommandation_SS_003_2011_0.pdf.

[89] Voir la délibération n° 14/059 du Comité sectoriel de la Sécurité Sociale et de la Santé, Section Santé, du 15 juillet 2014 relative à la communication de données à caractère personnel codées relatives à la santé dans le cadre du projet Thales, https://www.privacycommission.be/sites/privacycommission/files/documents/d%C3%A9lib%C3%A9ration_SS_059_2014.pdf.

caractère personnel que si ces mesures sont associées à la **garantie**[90] **qu'aucune tentative de réidentification ne sera entreprise**, par exemple au moyen d'un couplage avec d'autres ensembles de données.

✓ **Recommandation 7**

Le responsable du traitement doit prévoir l'agrégation ou l'anonymisation de données en tenant compte de la nature et du contexte du traitement. Si l'anonymisation de données est requise, l'admissibilité du risque d'individualisation ou d'identification indirecte dépend en effet de la nature de la partie qui utilisera les données et/ou de la finalité pour laquelle les données seront utilisées. Pour une réutilisation ou un traitement de données de santé par exemple, un niveau d'agrégation élevé sera dès lors souvent indiqué, tandis que pour d'autres applications (par exemple la lutte contre la fraude à l'énergie ou une recherche scientifique s'inscrivant dans un cadre réglementaire adéquat), on évaluera au cas par cas si un niveau d'agrégation plus faible peut se justifier.

D. "Données à caractère personnel exactes"

D'après la législation relative à la protection des données[91], la personne concernée a le droit d'obtenir sans délai du responsable du traitement la rectification des données à caractère personnel la concernant qui sont inexactes.

Dans le contexte du big data, la question se pose de savoir ce que signifie précisément la rectification de "données à caractère personnel inexactes". Il faut analyser ici aussi bien la dimension technique (mathématique – point D.1 ci-après) que la dimension sociale (Point E ci-après). Il est également utile de souligner la différence entre association par corrélation et association par causalité pour fournir une preuve (point D.2). La Commission formule ci-après quelques observations et recommandations concernant ces éléments.

D.1. Projections techniques optimales (les plus exactes possibles) quant aux personnes physiques

Obtenir des projections techniques optimales quant aux personnes physiques implique par exemple que le responsable doit veiller à ce que les modèles mathématiques utilisés soient "entraînés" de manière optimale, selon les règles de l'art, que le modèle procède donc au mieux à une généralisation de données de test indépendantes et représentatives (voir également le schéma 1 au point G) de sorte que la précision des projections soit optimale[92], que les variables (features) et ensembles de données (éviter une distorsion de données) soient choisis de manière optimale, que les résultats (corrélations) soient mis en œuvre correctement et que la précision des projections soit bien évaluée.

---

[90] Des engagements contractuels et (la possibilité de) contrôles via des audits peuvent, selon la situation, constituer des garanties adéquates.

[91] Voir l'article 12 de la LVP et l'article 16 du RGPD.

[92] Voir le point 35 de l'avis n° 25/2016 du 8 juin 2016 *relatif au projet de décret modifiant le Décret sur l'Énergie du 8 mai 2009, en ce qui concerne la prévention, la détection, la constatation et la sanction de la fraude à l'énergie.*

Concrètement, cela signifie notamment ce qui suit :

- ✓ Utilisation d'un ensemble d'apprentissage qui est représentatif pour la population sur laquelle on veut utiliser le modèle pour réaliser des projections (c'est généralement le cas si l'ensemble d'apprentissage est un échantillon aléatoire de cette population – cela signifie donc que l'ensemble d'apprentissage doit suivre la même distribution statistique que celle de la population ; voir également ci-après au sujet de la distorsion de données).
- ✓ Utilisation de techniques qui limitent, lors de l'apprentissage, le surapprentissage[93] du modèle (overfitting en anglais, c'est-à-dire la modélisation d'une variation aléatoire dans l'ensemble d'apprentissage qui n'a rien à voir avec la relation sous-jacente que l'on veut capter, par exemple lors de l'utilisation d'un trop grand nombre de variables par rapport au nombre de points ou d'observations)[94]. Les garanties nécessaires doivent donc être prises afin que le modèle généralise au mieux des données de test indépendantes de manière à ce que la précision des projections soit optimale.
- ✓ Utilisation d'un ensemble de test qui est également représentatif pour la population pour laquelle on veut utiliser le modèle pour réaliser des projections, ce qui est important pour effectuer une estimation exacte de la précision du modèle (voir aussi ci-après en ce qui concerne la distorsion de données).
- ✓ Utilisation d'un ensemble de test indépendant, c'est-à-dire qu'il ne peut en aucune manière avoir été utilisé lors de l'apprentissage du modèle (si cela se produit quand même, cela donnera lieu de manière générale à une surestimation de la performance du modèle).
- ✓ Étant donné que la population pour laquelle on veut réaliser des projections est dynamique dans de nombreux cas (à savoir que ses caractéristiques changent au fil du temps), il est important de procéder à intervalles réguliers à un réapprentissage des modèles avec un nouvel ensemble d'apprentissage représentatif du moment. Ensuite, l'évaluation du modèle doit également se faire avec un nouvel ensemble de test représentatif du moment.

- **Choix des ensembles de données et prévention de la distorsion de données**

D'après la littérature[95], le "big data" est souvent synonyme du fait de disposer de "plus de données". Le risque qui se présente ici est que pour réaliser des projections, on utilise des données désuètes, biaisées, non pertinentes ou non représentatives (pour la population pour laquelle on veut réaliser des projections)[96]. La situation est différente pour les données avec beaucoup de

[93] Voir le lexique.
[94] Pour une explication relative au supervised learning et au model training : voir notamment la partie I et le schéma 1.
[95] V. Mayer-Schönberger & K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work and think, Boston:Houghton Mifflin Harcourt, 2013.
[96] Voir les renvois à une mauvaise sélection des données, aux données incomplètes, incorrectes ou désuètes, aux préjugés lors de la sélection de la population, et à la confirmation et à la promotion involontaire de préjugés historiques aux pages 6 et 7 de Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, publié à l'adresse https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

parasites, qui permettent quand même souvent de réaliser des projections précises grâce à la taille des ensembles de données[97].

Des risques graves pour les droits et libertés des personnes concernées se présentent lorsqu'il est question de décisions liées à des modèles prédictifs de big data intégrant des préjugés (l'apprentissage se fait par exemple à l'aide de données qui ne sont pas représentatives pour la population pour laquelle on veut réaliser des projections, voir "distorsion de données")[98]. Un modèle prédictif de big data qui a été entraîné sur des données biaisées implique également une augmentation du risque que des personnes soient classées erronément et en outre que certains groupes de personnes soient systématiquement intégrés erronément à une classe déterminée (par exemple celle de "fraudeur potentiel").
Lors de la prévention de distorsion de données pour des données de test, le risque existe que la précision du modèle soit mal évaluée.

✓ Recommandation 8

Vu que le risque précité de "distorsion de données" peut se manifester lors de l'application du big data, les responsables (par exemple dans leurs analyses d'impact relatives à la protection des données) devront toujours **vérifier si une distorsion de données se dissimule dans les ensembles de données choisis et les méthodes pour procéder à l'apprentissage ou au test par exemple des modèles mathématiques**. Il faudra expliquer des questions telles que celles de savoir pourquoi certains groupes de population sont exclus, quels points de données sont moins visibles lors de l'apprentissage ou du test, …

- **Malentendus et conditions connexes pour de bonnes analyses de big data**

Il existe un malentendu selon lequel des algorithmes ou analyses de big data généreraient toujours des résultats parfaits (par exemple, des classifications sans erreur). Il ne faut pas s'imaginer que des erreurs pourraient être évitées du fait que l'on travaille avec de grands groupes de points de données. Même si des modèles mathématiques étaient entraînés et testés selon toutes les règles de l'art, des personnes seront encore toujours mal classées dans la plupart des projets (la performance de modèles mathématiques n'est jamais parfaite).

En outre, la qualité des algorithmes dépend en réalité des données avec lesquelles ils fonctionnent[99]. Toutes les données ne contiennent par exemple

---

[97] VAN DER SLOOT, B. et VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 32, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf.
[98] Préjugés cachés qui se dissimulent dans les données (les données ne sont pas représentatives de la population considérée). Voir par exemple CRAWFORD, Kate, The hidden biases in big data., 1er avril 2013, Harvard Business review, publié à l'adresse https://hbr.org/2013/04/the-hidden-biases-in-big-data, cité notamment dans FTC, Big Data. A Tool for inclusion or exclusion ? Understanding the Issues, janvier 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
[99] BAROCAS, S. et SELBST, A.D., Big Data's Disparate Impact, California Law Review 671, 11 août 2016, publié à l'adresse

pas suffisamment d'informations au sujet de ce que l'on veut modéliser (ce qui donne lieu à l'obtention de modèles dont la performance est faible ou inexistante). Le besoin d'un apport de données suffisant et correct est dans une situation de conflit par rapport au fait qu'il ne soit pas toujours évident ni même permis de se procurer ou de (ré)utiliser les données pertinentes. Les autorités fiscales mettent ainsi plus fréquemment en place des "fishing expeditions" au moyen de sources de données les plus grandes possible mais pas toujours fiables, comme dans le cas de l'examen de pages de réseau sociaux, forums de discussion et sites de vente, ou dans le cas de l'analyse massive de l'ensemble du groupe des utilisateurs de cartes de paiements (étrangères)[100]. Le problème réside dans le fait que les informations au sujet de ces populations ne sont pas toujours fiables ni représentatives pour examiner l'ensemble de la population des contribuables et de ceux qui éludent l'impôt, en les considérant comme "fraudeurs potentiels".

Dans les modèles prédictifs, on remarque aussi un glissement, dans l'appréciation des personnes, d'un modèle basé sur la causalité vers un modèle basé sur la corrélation[101]. Les corrélations qui sont trouvées à l'aide du big data ne sont pas toujours non plus mises en œuvre correctement. Il peut arriver que ces résultats soient généralisés (alors qu'ils ne sont par exemple valables que pour la population d'où provient l'ensemble d'apprentissage) ou que dans la communication ou la popularisation ultérieures des résultats de recherche, on ne tienne plus compte de la différence entre causalité et corrélation.

Il est donc crucial de disposer d'une compréhension technique minimale et de pouvoir utiliser de manière socialement responsable les "corrélations", la "distorsion de données" ainsi que la performance ou la puissance (et dans certain cas leur absence) de modèles. Sans cette compréhension, on peut parfois involontairement causer des effets sociaux et éthiques en utilisant des profils.

Il en résulte que le recours (technique) correct à l'utilisation d'analyses de big data n'est pas évident et est assorti d'innombrables conditions connexes. Il est dès lors préférable de prévoir à ce sujet une large obligation de minutie.

✓ Recommandation 9

Il convient de prévoir une **obligation d'attention/de minutie** à l'égard des responsables du traitement en ce qui concerne la qualité des données et la légitimité des méthodes d'analyse.

---

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[100] Fiscus screent massaal bankkaarten, De Tijd, 10 septembre 2016 ; Fiscus wil via bankkaarten jacht maken op verborgen buitenlandse rekeningen, http://deredactie.be/cm/vrtnieuws/binnenland/1.2763572.

[101] FTC, Big Data. A Tool for inclusion or exclusion ? Understanding the Issues, janvier 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

✓ **Recommandation 10**

Pour recourir à des analyses de big data en connaissance de cause, il est impératif que les contrôleurs, responsables et décideurs renforcent leur expertise en matière de big data et prévoient suffisamment de **formations** sur une base régulière pour recourir judicieusement et de manière responsable à ces techniques (ou à leur résultat).

D.2. Association par algorithme vs preuve juridique

L'utilisation de comparaison de fichiers, d'une statistique ou d'un algorithme qui démontre une association (relation) entre propriétés et phénomènes (par exemple relation entre consommation d'énergie inhabituelle et risque de fraude, relation entre retard de paiement en matière de téléphonie mobile et retard de paiement en matière de crédit[102], …) n'est pas d'emblée équivalente à la fourniture d'une preuve juridique d'une caractéristique d'une personne spécifique ou d'une relation de cause à effet entre deux phénomènes. Il subsiste toutefois souvent une méprise selon laquelle des relations de corrélation constituent une causalité (voir ci-avant), et que les relations ou projections de corrélation d'un modèle mathématique (qui peuvent donc être inexactes dans des cas individuels) impliquent une nécessité sociale urgente de lutter directement contre des phénomènes indésirables ou des "signalements" individuels dans le cadre de prévisions par le biais d'analyses de big data, comme dans le cas de la fraude sociale et des retards de paiement. Lors de l'utilisation d'analyses et de corrélations big data, il peut tout au plus être question d'un "clignotant" ou d'une indication[103] à laquelle le législateur ne peut pas appliquer attribuer une force probante (jusqu'à preuve du contraire par la personne concernée), mais uniquement une valeur d'indicateur qui doit être analysé plus avant dans une enquête individuelle et personnalisée, par le biais notamment d'une intervention par une personne physique. Dans le cadre par exemple de la détection de la fraude, de telles analyses servent uniquement à mettre en œuvre le plus efficacement possible (avec un résultat maximalisé) les moyens limités du fait que l'on peut diriger l'enquête vers un groupe de personnes ou d'organisations présentant un risque accru de fraude (au lieu de procéder par exemple à des contrôles non ciblés et basés sur des échantillons de la population générale), sans nécessairement qu'il n'existe pour autant déjà des preuves concluantes au niveau de l'analyse de données. En d'autres termes, l'objectif est donc de réduire le spectre en ciblant davantage. Il faut toutefois tenir compte de la marge d'erreur (la possibilité d'erreur de classification d'individus par des modèles) et de l'impact social de telles mesures et prévoir des garanties pour protéger suffisamment les personnes concernées.

Partir (erronément) du principe que des corrélations de fraude, de retards de

[102]Voir le document de travail n° 112 de la Banque nationale de 2011 qui a démontré une relation entre retards de paiement en matière de téléphonie mobile et retards de paiement en matière de crédit : https://www.nbb.be/en/articles/working-paper-ndeg-212.

[103] Voir la loi du 13 mai 2016 modifiant la loi-programme (I) du 29 mars 2012 *concernant le contrôle de l'abus d'adresses fictives par les bénéficiaires de prestations sociales, en vue d'introduire la transmission systématique de certaines données de consommation de sociétés de distribution et de gestionnaire de réseaux de distribution vers la BCSS améliorant le datamining et le datamatching dans la lutte contre la fraude sociale*, M.B., 27 mai 2016.

paiement, … impliquent toujours une vérité objective sur la base de laquelle il faut toujours prendre des mesures, c'est un phénomène que l'on appelle aussi "**data determinism**"[104].

✓ **Recommandation 11**

Il est important de prévoir des règlements légaux qui déterminent comment et quand le résultat du data mining et des analyses statistiques (corrélations) peut être utilisé par les autorités en tant que preuve juridique.

✓ **Recommandation 12**

Pour les applications tant dans le secteur public que privé (par exemple pour l'octroi de crédits, les recrutements, …), il faut donc prévoir suffisamment de transparence et de **garanties de procédure** qui assurent un mode de décision correct à l'égard des individus, lequel dépasse une intervention basée uniquement sur une corrélation ou une relation statistique.

Un exemple d'une telle garantie dans la lutte contre la fraude est le fait de prévoir une obligation légale de dresser un procès-verbal de constatation[105] reprenant les constatations physiques ou l'analyse personnalisée d'un enquêteur ainsi qu'une obligation de signification du procès-verbal avec un temps de réaction pour la personne concernée. En outre, personne ne peut dans ce cas être formellement accusé ou sanctionné sur la base de l'utilisation de relations statistiques, du résultat de modèles mathématiques ou de comparaison de fichiers. Cela peut toutefois justifier une enquête plus approfondie.

E. Impact individuel ou collectif pour les droits et libertés des personnes concernées

Les projets de big data peuvent avoir un impact sur les droits et libertés de personnes physiques (surtout si des choix techniques non optimaux ont été faits, mais même aussi si on suit les règles de l'art). Le résultat peut aussi ne pas être socialement acceptable, par exemple parce que des dispositions de discrimination ne sont pas respectées (point E.2). Dans les deux cas, le principe de traitement loyal de données à caractère personnel peut être enfreint (voir aussi ci-après au point F).

E.1. Impact du big data sur les personnes physiques individuelles

Le RGPD[106] requiert, comme déjà mentionné ci-avant, que le responsable effectue une évaluation continue des risques et une analyse d'impact relative à la protection des données en fonction de l'impact sur les droits et libertés des

[104] Data determinism – can data really speak for itself?, 5 février 2014, https://datasciencebusiness.wordpress.com/2014/02/05/data-determinism/.
[105] Voir le renvoi aux points 39 e.s. de l'avis n° 24/2015 du 17 juin 2015 relatif à la fraude sociale (référence aux garanties) dans la jurisprudence fiscale concernant le droit de visite fiscal.

personnes concernées.

Une application approfondie de projections à l'égard de personnes individuelles entraîne un risque d'ingérence grave dans les droits et libertés de cet individu. Cela peut être le cas pour les contrôles réalisés par les services de police, basés sur une analyse de profils Facebook (predictive policing), l'utilisation de corrélations entre consommation d'énergie faible ou élevée et la prévention de fraude (perte d'une allocation), ou l'utilisation de données via des applications smartphone par les assureurs pour la détermination d'une prime d'une assurance maladie ou l'assurance responsabilité civile pour véhicules à moteur[107].

E.2. Impact social du big data sur certains groupes sociaux

On a déjà évoqué ailleurs[108] l'utilisation de certaines corrélations, ensembles de données ou algorithmes qui produisent une stratification sociale (ce qu'on appelle placer une population dans des catégories de "social sorting" [109]), impliquant un traitement potentiellement illégal et inégal de groupes sociaux. Il est certes possible que ce traitement inégal ne soit pas intentionnel mais se dissimule dans les données ou les algorithmes.

Tous les traitements inégaux de personnes ne sont toutefois pas illégaux [110] . C'est toutefois le cas s'il est par exemple question de discrimination[111].

Au niveau technique, les mécanismes suivants[112] peuvent se produire dans ce cadre :

- ✓ **Définition des étiquettes de classification qui placent les membres d'un groupe social de manière préférentielle dans une catégorie qui est défavorisée** (par exemple lorsque dans le cadre d'une décision de recrutement, on utilise un modèle qui fait une distinction entre de "bons" ou de "mauvais" travailleurs, la distinction étant faite en comparant la durée de la période de service du travailleur avec une valeur arbitraire). Lorsque les membres de certains groupes sociaux changent plus souvent de travail pour quelque raison que ce soit, cela réduira leurs chances d'être engagés (du fait de l'utilisation d'un

[107] VAN DER SLOOT, B. et VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p. 100, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf.

[108] Voir le système d'évaluations de crédits aux USA, cité aux pages 11 à 13 inclus de Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Mei 2016, publié à l'adresse https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf ; Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, point 4.5.1., page 89.

[109] D. Lyon, Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination, London, Routledge, 2003, disponible à l'adresse https://infodocks.files.wordpress.com/2015/01/david_lyon_surveillance_as_social_sorting.pdf et BROEDERS, D. et SCHRIJVERS, E., *Big Data in een vrije en veilige samenleving*, 20 mai 2016, http://njb.nl/njv-jaarvergaderingen/jaarvergadering-2016/artikelen/big-data-in-een-vrije-en-veilige-samenleving.19799.lynkx.

[110] Cela n'empêche toutefois pas qu'un traitement légal mais inégal puisse comporter des risques pour les droits et libertés des personnes physiques concernées, ce dont le responsable du traitement devra tenir compte.

[111] Les lois anti-discrimination et antiracisme ont un volet tant civil que pénal. Pour le volet pénal, une intention particulière est requise mais pour le volet civil, l'intention n'a aucune importance.

[112] Solon Barocas and Andrew D. Selbst, (2016) Big Data's Disparate Impact, 104 Cal. L. Rev. 671, publié à l'adresse http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

modèle entraîné sur des données dotées de telles étiquettes de classification), même s'ils sont plus productifs que d'autres.

- ✓ **Données d'apprentissage qui sont biaisées d'une manière ou d'une autre**. Par exemple parce qu'elles sont basées sur des données d'apprentissage où **l'attribution d'une étiquette de classification est déjà basée sur un préjugé**. Dans ce cas, le modèle qui en résulte ne fera que perpétuer le préjugé. Ou par exemple parce qu'il est basé sur **des données d'apprentissage où on a choisi de préférence dans un certain groupe de population des personnes qui se trouvent dans la "mauvaise" catégorie ou la catégorie la plus défavorable**.
- ✓ Lorsque l'appartenance à un groupe de population (ou un mandataire à cet égard ; voir ci-après au sujet des critères protégés) qui présente un moins bon score dans un certain domaine (par exemple le taux de criminalité plus élevé au sein d'un groupe ethnique déterminé) constitue un critère sur la base duquel un modèle prend une décision dans ce cadre, **l'utilisation d'un ensemble trop limité** – qui ne permet pas de modéliser la variabilité naturelle présente dans ces groupes de population – **de variables ou de caractéristiques** dans le modèle peut générer un modèle qui prend des décisions défavorables à l'égard de l'ensemble des membres d'un groupe déterminé (par exemple l'appartenance à un groupe ethnique indique en soi un risque bien plus élevé de prévision de comportement criminel, sans tenir compte des autres caractéristiques d'un individu).
- ✓ L'**utilisation intentionnelle** des mécanismes décrits ci-avant (qui se mettent généralement en place de manière non intentionnelle) dans le but de discriminer certains groupes.

Le **traitement inégal** de personnes physiques qui ne constitue pas une **discrimination interdite** au sens de la législation applicable[113] ne peut a priori pas être considéré comme illégal en vertu de la législation anti-discrimination.

Lorsque l'on aborde des problèmes sociaux à l'aide du big data, on peut voir apparaître, par des fausses corrélations[114], le "fondamentalisme de données"[115] et la "distorsion de données", des effets néfastes pour une partie spécifique de la population. Des enquêtes sur certains algorithmes concernant le transport[116] ou la circulation[117] ont démontré le risque qu'une répartition d'avantages basée sur un algorithme de décision puisse donner lieu à une répartition inégale de moyens (publics) ou de ces avantages (réparation de routes, prestation ou non

---

[113] Voir notamment la loi du 10 mai 2007 *tendant à lutter contre certaines formes de discriminations*, M.B., 30 mai 2007.

[114] Phénomène par lequel on suppose une relation entre un critère sensitif (par exemple l'ethnicité) et ce que l'on veut prédire, tandis que d'autres critères (moins sensitifs) permettent en soi de meilleures projections, sans avoir besoin de ce critère sensitif.

[115] Partir erronément du principe que la corrélation indique toujours une relation causale et que des ensembles de données massifs et des analyses prédictives donnent toujours la vérité objective. Voir CRAWFORD, Kate, The hidden biases in big data., 1er avril 2013, Harvard Business review, publié à l'adresse https://hbr.org/2013/04/the-hidden-biases-in-big-data et page 10 de Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, publié à l'adresse https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[116] Voir par exemple l'étude sur l'algorithme "surge pricing" de Uber, qui a démontré que cet algorithme générait des tarifs plus élevés et de meilleurs services pour certains quartiers, et des tarifs inférieurs et un mon bon service pour d'autres quartiers. DIAKOPOLOUS, N., How Uber surge pricing really works, 17 avril 2015, https://www.washingtonpost.com/news/wonk/wp/2015/04/17/how-uber-surge-pricing-really-works/.

[117]Voir la "Streetbump smartphone app" que la ville de Boston voulait utiliser pour allouer des moyens pour des travaux routiers. Cette méthode a toutefois été hantée par le préjugé dissimulé selon lequel chaque personne avait accès à cette app. Voir CRAWFORD, Kate, The hidden biases in big data., 1er avril 2013, Harvard Business review, publié à l'adresse https://hbr.org/2013/04/the-hidden-biases-in-big-data.

d'un service dans un temps de réaction donné, octroi d'une réduction sur des services). Il peut en résulter que des moyens ne soient pas toujours appliqués efficacement à l'égard de groupes de population en raison de suppositions incorrectes ou du fait qu'un traitement différent à l'égard de certains groupes de population est perpétué, aggravé ou masqué [118].

D'autres exemples sont la discrimination tarifaire et/ou les projets de big data qui compromettent l'accès égal à des opportunités sur le marché du travail ou la diversité sur le lieu de travail [119]. Par exemple, un bureau de ressources humaines qui, à l'aide d'un logiciel de recrutement, appliquerait aveuglément une stratégie "hiring for culture fit"[120] avec une discrimination inconsciente[121] sur la base de critères protégés tels que la religion ou les convictions politiques ou philosophiques, l'origine ethnique, l'âge, le sexe ou la race, la grossesse, la maternité et le handicap[122]. Dans cet exemple, il faudrait veiller à ce qu'un algorithme ne refuse pas la participation sociale à des groupes historiquement défavorisés ou vulnérables.

✓ Recommandation 13

Les **responsables du traitement** prennent les précautions nécessaires pour éviter au maximum que des analyses de big data donnent lieu à des formes interdites de discrimination[123] dans certains groupes de population ou à des traitements inégaux qui peuvent constituer un risque élevé pour les droits et libertés des personnes concernées. L'étude de l'impact sur certains groupes sociaux ou culturels peut se faire dans les analyses d'impact relatives à la protection des données.

Le **législateur** pourrait aussi recourir au principe "d'égalité des chances dès la conception" ("**equal opportunity by design")**[124], afin d'éviter que soient utilisées des techniques de big data qui sont contraires aux dispositions anti-discrimination ou au principe de loyauté (voir également ci-après le point F). À cet égard, on pourrait se concentrer sur des critères ou variables protégés, qui ne peuvent pas être utilisés, ou uniquement avec la plus grande prudence, dans des modèles qui prennent des décisions qui affectent grandement les personnes.

---

[118] Voir p. 12 de Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, publié à l'adresse https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[119] DE ANCA, C., Why hiring for cultural fit can thwart our diversity efforts, Harvard Business review, 25 avril 2016, https://hbr.org/2016/04/why-hiring-for-cultural-fit-can-thwart-your-diversity-efforts.

[120] Ce du postulat que le candidat souhaité présente de préférence une corrélation avec l'origine ethnique, le groupe d'âge ou la race dominants dans une entreprise. Voir HUHMAN, H., Not Using Big Data for Hiring? You May Be Missing Out on the Best Candidates, 10 April 2014, https://www.entrepreneur.com/article/232780.

[121] Les lois anti-discrimination et antiracisme ont un volet tant civil que pénal. Pour le volet pénal, une intention particulière est requise mais pour le volet civil, l'intention n'a aucune importance.

[122] Ce critère apparaît explicitement dans les législations respectives mais aussi dans la Directive 2000/78/UE et son importance augmente vu la Convention des Nations Unies relative aux droits des personnes handicapées. Depuis que l'Europe a signé cette Convention, la directive doit être mise en œuvre conformément à cette dernière.

[123] Voir notamment la loi du 10 mai 2007 *tendant à lutter contre certaines formes de discriminations*, M.B., 30 mai 2007.

[124] Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, publié à l'adresse https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Des corrections sociales dans des dossiers individuels peuvent constituer une solution partielle si l'on utilise potentiellement des modèles présentant une distorsion de données ou proportionnellement de mauvaises classifications dans certains groupes. Ces corrections peuvent prendre la forme d'une possibilité de recours, de plainte ou d'appel, une demande de révision de classification de données dans des dossiers, le fait de prévoir des mécanismes de feed-back qui tentent de réduire ultérieurement des classifications erronées dans de tels cas.

Au niveau collectif, la législation peut protéger les personnes concernées contre une réduction trop sévère d'opportunités en matière de services de base suite à un profilage trop avancé sur la base de critères ou de variables protégés (l'âge ou des données de santé par exemple). C'était par exemple le cas dans la loi Partyka[125] concernant la souscription à des assurances solde restant dû par des personnes présentant un risque de santé accru, loi qui a introduit un système de questionnaires approuvés.

✓ Recommandation 14

Des corrections sociales [126] au niveau individuel ou collectif doivent être disponibles afin d'éviter des préjugés au stade le plus précoce, de respecter les dispositions anti-discrimination (qui sont d'ordre public), et d'éviter en temps utile des inconvénients cumulatifs pour certains groupes de population (risque de discrimination ou de traitement de données déloyal dus à une différence de traitement).

✓ Recommandation 15

Il est important, du point de vue de l'obligation de motivation [127] et/ou de l'analyse d'impact relative à la protection des données [128] du RGPD, que les responsables prévoient des mécanismes de motivation qui font la transparence quant à la base sur laquelle l'algorithme prend des décisions (voir aussi ci-après au point L) et qui fait l'objet d'un feed-back individuel et/ou social, et ce surtout si les décisions de l'algorithme peuvent toucher fortement les personnes concernées.

Le feed-back se révélera nécessaire s'il y a une combinaison de plusieurs risques élevés pour les droits et libertés des personnes concernées, d'erreurs ou d'incidents de sécurité [129] concernant des formes extrêmes d'analyses de big

---

[125] La "loi Partyka" concerne les articles 212 à 224 de la loi du 4 avril 2014 *relative aux assurances* (Moniteur belge du 30 avril 2014, publiée au M.B. du 3 février 2010. Les arrêtés d'exécution sont parus avec l'arrêté royal du 10 avril 2014 *réglementant certains contrats d'assurance visant à garantir le remboursement du capital d'un crédit hypothécaire* (M.B. du 10 juin 2014).

[126] Via par exemple des dispositions dans la législation qui se penchent sur le risque de discrimination ou de traitement inégal de data mining.

[127] Article 5.2 du RGPD.

[128] Article 35.9 du RGPD

[129] STROUD, M., The minority report: Chicago's new police computer predicts crimes, but is it racist? Chicago police say its computers can tell who will be a violent criminal, but critics say it's nothing more than racial profiling, The Verge, 19 février 2014, http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist.

data (c'est-à-dire avec un risque résiduel trop élevé).

Selon le contexte, ce feed-back[130] peut prendre diverses formes, comme l'intégration d'une "modélisation interactive"[131] (où les personnes concernées (clients) peuvent contribuer à la détermination de l'algorithme), un mécanisme de filtre[132] (par exemple des mesures qui pourraient limiter la diffusion de messages erronés via les réseaux sociaux), le fait de prévoir une transparence sociale suffisante et un moment d'évaluation formel et/ou une obligation de rapport particulière à l'égard des contrôleurs et de la population quant à l'impact du modèle.
Il convient toutefois dans ce cadre de veiller à ce que des garanties adéquates soient offertes pour trouver un équilibre entre la mise à disposition d'un feed-back et la protection des intérêts commerciaux ou généraux des responsables du traitement ou la sécurité de traitements[133].

Pour des applications concrètes, cet équilibre peut également être prévu par l'examen critique et indépendant du fonctionnement d'algorithmes, de comparaisons de fichiers, … et ce dans un contexte social plus large.

✓ **Recommandation 16**

Les modèles ne sont jamais parfaits et les erreurs de classification peuvent être concentrées dans certains groupes de population. Si des problèmes peuvent se poser dans ce contexte, l'output d'algorithmes doit être analysé de manière critique et si possible indépendante au niveau des erreurs et le modèle doit le cas échéant être adapté ou corrigé.

F. Principe de loyauté

Le principe de loyauté qui est actuellement repris à l'article 4, § 1, 1° de la LVP et à l'article 5.1 (a) du RGPD[134] est en lien direct avec l'exigence de traitement transparent et le mode d'obtention des données.

La Commission distingue plusieurs hypothèses dans lesquelles le principe de loyauté peut être enfreint.

D'une part, il peut être question de l'absence de traitement transparent[135] (voir aussi ci-après au Point L). D'autre part, l'application d'un modèle peut impliquer

[130] VEALE, M., Shadow of the smart machine: Computer scientists and social scientists must work together on algorithms, 26 janvier 2016, Shadow of the smart machine series, http://www.nesta.org.uk/blog/shadow-smart-machine-computer-scientists-and-social-scientists-must-work-together-algorithms.
[131] Voir une application dans le contexte du journalisme : la conclusion de l'article de DIAKOPOULOS, N., Accountability in Algorithmic Decision Making, février 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/2016/03/Accountability-in-algorithmic-decision-making-Final.pdf.
[132] SOLON, O., Facebook's failure: did fake news and polarized politics get Trump elected?, 10 novembre 2016, https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories.
[133] Considérant 63 et article 35.9 du RGPD.
[134] Le terme anglais "fairness" a été erronément traduit en néerlandais par "behoorlijkheid", et correctement en français par "loyauté".
[135] Les entreprises qui, sans information suffisante des personnes concernées, collectent des données ou effectuent des corrélations et les répartissent (par exemple via identification à l'aide d'un couplage avec des données ouvertes) ne respectent pas le principe de loyauté.

la violation de diverses dispositions d'interdiction en matière de discrimination[136] au niveau régional, fédéral ou européen (voir également ci-avant au point F).

G. Proportionnalité : nécessité, mode de stockage et principe de traitement de données minimal

Une ingérence dans la vie privée n'est légitime que si l'on peut démontrer que celle-ci **est nécessaire dans une société démocratique**[137]. Les données utilisées doivent en outre être non excessives et adéquates à l'égard de la finalité pour laquelle elles sont traitées.

Vu le risque de nombreux pièges (distorsion de données, faux positifs et faux négatifs, data determinism, …), les analyses de big data ne sont a priori jamais une méthode efficace de prévoir des comportements individuels avec précision. L'efficacité de cette méthode dépend de la question de savoir si on dispose de suffisamment de données pertinentes et informatives au sujet du phénomène à étudier et/ou si on peut déduire des données des schémas significatifs à l'aide de la bonne méthode, et si les citoyens adaptent leur comportement en réaction à des pratiques d'observation à grande échelle. Il a ainsi été affirmé[138] que le data mining pourrait être une méthode inefficace dans la recherche de terroristes, de la criminalité organisée ou du trafic des êtres humains étant donné que ces phénomènes présentent souvent un caractère insuffisamment régulier, qu'il y a trop peu de données fiables ou suffisantes, et qu'il y a trop peu de conventions que pour établir un bon profil.

Malgré les promesses d'obtenir des résultats positifs lors de la mise en œuvre de big data pour lutter contre la fraude et certaines formes de criminalité, il manque dans de nombreux cas un fondement fiable et objectif et une évaluation de l'efficacité des méthodes d'analyse utilisées.

✓ Recommandation 17

L'utilisation de big data dans le secteur public, en particulier pour la promotion de la sécurité nationale, la lutte contre la criminalité et le maintien de l'ordre, doit toujours être soumise à une enquête des contrôleurs au niveau de l'opportunité et de l'efficacité[139].

Un autre risque à la lumière du principe de proportionnalité réside dans le fait que l'on compare trop vite les possibilités et conditions des méthodes de surveillance classiques de l'observation (policière) de l'espace public avec l'observation en ligne. Les "patrouilles" numériques de la police (par exemple sur des sites Internet, des forums, des réseaux sociaux) ou des formes de

---

[136] Pour un aperçu de la législation pertinente, voir la page 7 du "Lexique discrimination" de UNIA, publié à l'adresse http://www.unia.be/fr/legislation-et-recommandations/legislation/lexique-discrimination.

[137] Cette exigence ne découle pas du RGPD, mais de l'article 8 de la CEDH et de l'article 22 de la Constitution. Voir également les articles 7 et 52 de la Charte européenne.

[138] Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, p. 83 point 4.3.2.

[139] VAN DER SLOOT, B. et VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 28, publié à l'adresse http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf.

"predictive policing", comme dans le projet i-Police récemment annoncé[140], se font à une autre échelle et à l'aide d'autres méthodes (stockage et analyse de données) que celles disponibles pour les observations proactives "hors ligne" des lieux accessibles au public à des fins de police administrative ou judiciaire.

Une modification de la loi sur la fonction de police[141] qui avait été prévue avait tenté d'assimiler les patrouilles numériques à la pénétration par la police de lieux accessibles au public. La Commission avait toutefois associé [142] ces patrouilles numériques aux méthodes particulières de recherche, et a également pointé des problèmes par rapport au principe de légalité.

Il découle par ailleurs du droit relatif à la protection des données que les données à caractère personnel ne peuvent pas être conservées pour une durée excédant celle nécessaire à la réalisation des finalités pour lesquelles elles sont traitées (sous une forme permettant l'identification des personnes concernées) (principe de limitation du stockage[143]). Dans le cadre de projets de big data, où les données sont souvent stockées pour une durée indéterminée en attente d'une application utile (traitement ultérieur – voir également le point H), il peut être question d'une ignorance de ce principe de limitation.

L'utilisation d'une quantité de données à caractère personnel supérieure à celle nécessaire à une finalité déterminée (ce qui est souvent le cas dans les projets de big data) va à l'encontre du "**principe de traitement de données minimal**"[144]. On a notamment tenu compte de ce principe pour l'archivage de données (à caractère personnel) où les fichiers d'archive de l'UE sont généralement soumis à des critères de sélection stricts et mûrement réfléchis et dont seule une fraction est conservée, au lieu de redéfinir le rôle de l'archiviste en tant que créateur proactif de possibilités d'autoriser la réutilisation de données pour d'autres motifs. Si des données à caractère personnel sont traitées dans ce cadre, il faut toutefois veiller à ce que l'approche dans d'autres pays (par exemple le projet Twitter Research Access aux USA[145]) ne disposant pas d'un règlement de protection des données similaire à celui de l'UE ne soit pas reprise aveuglément.

La Commission énumère ci-après quelques critères à l'aide desquels la proportionnalité du big data (et de projets d' "open data", voir ci-après) peut être évaluée.

a) Pour évaluer la proportionnalité par exemple de projets commerciaux de big data ou d'open data ou de projets d'archivage de services publics, il peut être important de vérifier s'il est ou non question d'une phase d'agrégation suffisante ou d'anonymisation. Une analyse de risques

---

[140] Ponciau, L, Les détails du projet iPolice dévoilés, Le Soir, 17 septembre 2016.
[141] Voir les points 14 et suivants de l'avis n° 13/2015 du 13 mai 2015 *relatif aux avant-projets de loi portant dispositions diverses – modifications de la loi portant création d'un organe de recours en matière d'habilitations de sécurité, de la loi sur la fonction de police et de la loi du 18 mars 2014 relative à la gestion de l'information policière*, publié à l'adresse https://www.privacycommission.be/sites/privacycommission/files/documents/avis_13_2015.pdf.
[142] Voir le point 32 de l'avis précité.
[143] Article 5.1. (e) du RGPD.
[144] Article 5.1. (c) du RGPD
[145] Voir ADAMS, R., All your Twitter belongs to the Library of Congress, The Guardian, 14 avril 2010, publié à l'adresse https://www.theguardian.com/world/richard-adams-blog/2010/apr/14/twitter-library-of-congress ; SCOLA, N., Library of Congress' Twitter archive is a huge #FAIL. More than five years on, the library's Twitter archive project is in limbo — with no end in sight, 7 novembre 2015, publié à l'adresse http://www.politico.com/story/2015/07/library-of-congress-twitter-archive-119698.html#ixzz4KJaJhExE.

"Small Cells" (voir à ce sujet le point C) peut être utile dans ce cadre. Des cas dépourvus de phase d'agrégation peuvent ainsi être considérés plus rapidement comme étant disproportionnés.

b) Le principe de proportionnalité signifie que lors des traitements, il faut également être attentif à la fréquence de lecture[146] des données (à caractère personnel) en fonction des fonctionnalités nécessaires, par exemple pour le traitement de données de compteurs (d'énergie) intelligents ou d'applications de smartphones.

c) On retrouve un autre critère d'évaluation de la proportionnalité de traitements (publics) dans la jurisprudence de la Cour européenne des droits de l'homme (CEDH)[147]. Lors de l'étude de proportionnalité, la Cour s'intéresse à la possibilité de mettre en œuvre des méthodes de recherche alternatives et moins poussées au lieu d'opter pour une "surveillance totale et universelle". On peut également se référer à cette jurisprudence pour l'application de techniques de data mining dans le cadre des patrouilles numériques (voir ci-avant), de la "predictive policing" et de la détection de la fraude à l'aide de corrélations (clignotants). Même dans la lutte contre la fraude fiscale, la fraude sociale, la fraude à l'énergie, … la proportionnalité des traitements de données doit être prise en compte et il faut vérifier si les mêmes opportunités sociales ne peuvent pas être réalisées au moyen de méthodes de moins grande ampleur.

d) Des modèles tels que le "predictive policing" et la détection de la fraude à l'aide de corrélations (clignotants) comme dans la lutte contre la fraude fiscale, la fraude sociale, la fraude à l'énergie, la fraude en matière de paiements, … dans les secteurs public et privé doivent être traités par le législateur comme des **méthodes de recherche extraordinaires**[148], et doivent disposer d'un cadre légal similaire qui fixe suffisamment les limites de ces méthodes et les garantit. À défaut d'un **cadre légal suffisamment clair**[149] pour les modèles précités, le respect du principe de proportionnalité pour ces modèles ne peut pas être vérifié correctement et il n'y a aucune garantie que la technique de recherche particulière est mise en œuvre de manière socialement responsable. Il est toutefois important que ces garanties légales soient formulées de manière générique et neutre du point de vue technologique (donc sans référence à des techniques spécifiques et à une méthode de data mining qui peuvent être actuelles ou s'appliquer à un moment déterminé).

e) Lors de l'élaboration de modèles prédictifs, il est préférable de travailler en deux phases :

---

[146] Voir la page 15 de l'avis du Régulateur flamand du marché de l'électricité et du gaz (VREG) du 8 avril 2015 sur un projet d'arrêté du Gouvernement flamand modifiant l'arrêté relatif à l'énergie du 19 novembre 2010, en ce qui concerne l'installation de compteurs intelligents, publié à l'adresse http://www.vreg.be/sites/default/files/document/adv-2015-03_ontwerp_van_besluit_uitrol_slimme_meters.pdf.

[147] CEDH, 2 septembre 2010, Uzun c. Allemagne.

[148] Voir le point 14 de l'avis 25/2016 du 8 juin 2016 concernant le *projet de décret modifiant le Décret sur l'Énergie du 8 mai 2009, en ce qui concerne la prévention, la détection, la constatation et la sanction de la fraude à l'énergie*.

[149] Référence à la loi au sens matériel et large, cela comprend donc aussi d'éventuelles autorisations émises par des comités sectoriels, lesquelles peuvent comporter une étude de proportionnalité de la demande.

✓ Recommandation 18

En cas d'élaboration et d'utilisation de modèles prédictifs ((semi-) supervised learning), il est préférable de travailler en deux phases (voir le schéma 1). Dans la première phase ou la phase préalable, on peut établir un modèle mathématique et procéder à son apprentissage à l'aide de données les plus anonymes possible où l'on a au moins supprimé les identifiants directs. Par ailleurs, on peut utiliser au cours de cette phase toutes les données ou variables (y compris les étiquettes de classification) dont on dispose (et bien entendu à condition qu'elles soient traitées de manière licite), les stocker et les prendre en considération pour l'apprentissage ou l'éventuelle inclusion dans le modèle final. Dans une deuxième phase (qui concerne le "singling out" et/ou l'identification de la personne concernée[150]) de l'application opérationnelle du modèle entraîné au cours de la phase précédente, on ne peut utiliser, collecter et enregistrer que les données/variables qui, au cours de la phase d'entraînement, ont significativement contribué aux projections du modèle (principe de traitement de données minimal).

On peut également travailler en deux phases en introduisant notamment des rôles distincts : analystes (responsables de l'entraînement et du test des modèles) et opérateurs (responsables de l'utilisation de modèles opérationnels dans les processus décisionnels). Cela signifie donc que les analystes ont accès à toutes les variables qui sont disponibles mais qu'ils ne peuvent pas être en mesure de réidentifier les points de données. Les opérateurs connaissent par contre bel et bien l'identité derrière les points de données mais n'ont accès qu'à un nombre limité de variables qui sont incluses dans les modèles opérationnels.

[150] Voir aussi le point 35 de l'avis 25/2016 du 8 juin 2016.

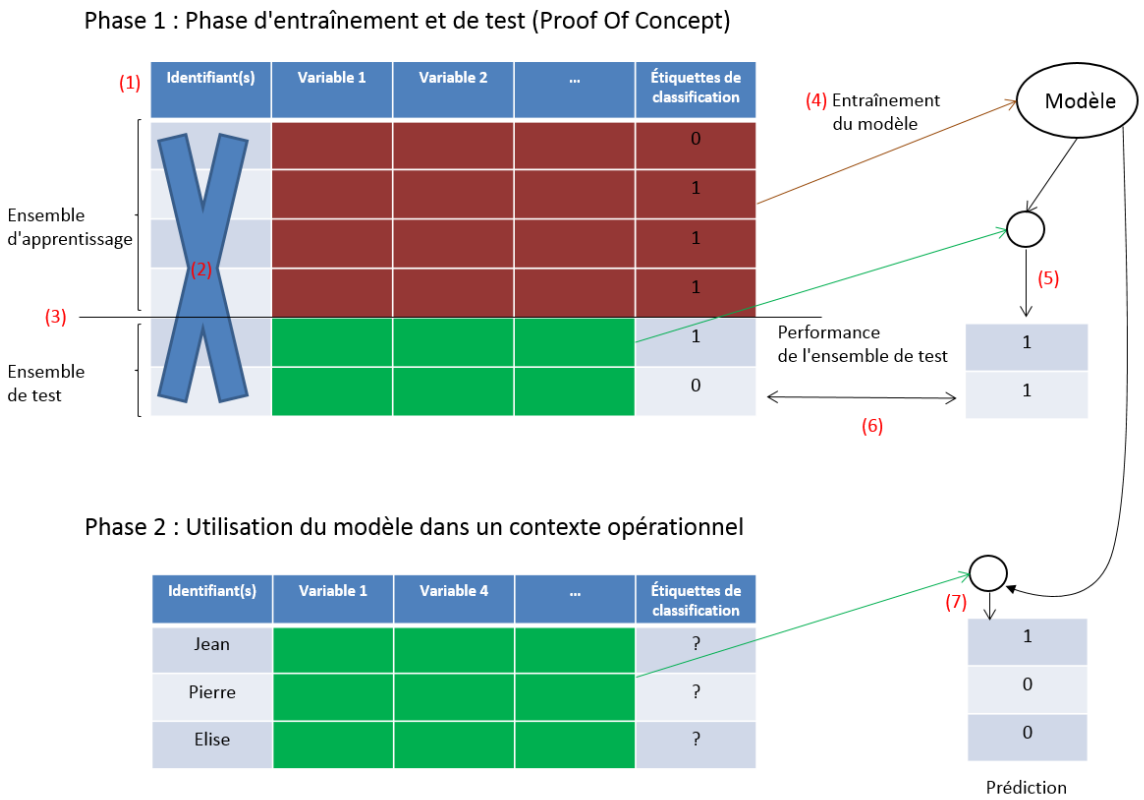


Schéma 1
Predictive modelling : différentes phases :

Phase 1 : Phase d'entraînement et de test

(1) On part d'un ensemble de données où, pour chaque individu (les lignes), on connaît tant les variables d'input que la (classe de) catégorie à laquelle l'individu appartient (par exemple solvable ou non, présence ou non d'une maladie, intéressé ou non par un certain produit, … ; dans l'exemple, on travaille avec deux catégories présentées par 1 ou 0. C'est ce qu'on appelle aussi les étiquettes de classification). L'objectif dans cette phase est d'établir et de tester un modèle qui peut prédire la catégorie d'un individu à l'aide des variables d'input. À cet effet, il n'est pas nécessaire de connaître l'identité des personnes dans l'ensemble de données.

(2) Suppression des identifiants ou utilisation d'autres techniques d'anonymisation. Lors de cette première phase, on ne peut pas tenter de réidentifier les individus à l'aide des variables subsistantes après anonymisation.

(3) Scission de l'ensemble de données en un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage sera utilisé pour établir le modèle et l'ensemble de test sera ensuite utilisé pour mesurer la précision du modèle. Les données de l'ensemble de test ne peuvent en aucune façon être utilisées pour l'établissement ou l'apprentissage du modèle (ensemble de test indépendant). Tant l'ensemble d'apprentissage que l'ensemble de test doivent être représentatifs de la population dans la Phase 2 pour laquelle on veut utiliser le modèle pour réaliser des projections.

(4) Apprentissage : établissement du modèle à l'aide des données de l'ensemble d'apprentissage.

(5) Test : le modèle est appliqué aux individus de l'ensemble de test. Projection des étiquettes de classification des personnes de l'ensemble de test.

(6) Comparaison de la valeur prédite des étiquettes de classification avec leur valeur réelle. On peut ainsi calculer la précision du modèle ou la performance de l'ensemble de test indépendant.

Phase 2 : Utilisation du modèle dans un contexte opérationnel sur des données de nouvelles personnes (identifiées) où la catégorie à laquelle les individus appartiennent n'est pas encore connue. Ici aussi, on ne peut accéder qu'aux données/variables qui, au cours de la phase d'apprentissage, ont contribué significativement aux projections du modèle.

(7) Le modèle est utilisé pour prédire les étiquettes de classification pour de nouvelles personnes et ces informations sont utilisées pour déterminer de futures décisions quant à cet individu.

H. Principe de finalité

Le droit relatif à la protection des données comporte l'obligation de ne pas utiliser les données pour des finalités autres que celles pour lesquelles les données ont été obtenues initialement ("exigence d'utilisation compatible"). Ce principe est mis à mal dans de nombreuses applications de big data. Les projets de big data concernent en effet souvent une utilisation de données pour une finalité autre que celle pour laquelle les données ont été collectées initialement (on procède d'abord à la collecte et on analyse ensuite ce qu'on peut en faire), parfois aussi après couplage avec des données provenant d'autres sources. Une telle réutilisation peut souvent créer une grande plus-value – souvent aussi au niveau social –, comme par exemple pour détecter des affections médicales rares ou dans la lutte contre la fraude fiscale ou la fraude sociale, mais est cependant contraire à l'exigence d'utilisation compatible.

Quels facteurs peuvent être importants pour considérer un traitement ultérieur comme étant compatible avec le traitement initial ? L'article 6.4 du RGPD comporte une liste non limitative de facteurs pour évaluer cette question :

✓ Recommandation 19

La Commission recommande aux responsables de tenir compte des éléments suivants lors de l'évaluation de compatibilité de traitements.

a) Le lien entre les finalités pour lesquelles les données ont été traitées et les finalités qui résultent de l'analyse de big data. La prévision d'un certain résultat n'est pas toujours en adéquation avec la finalité initiale. Des études scientifiques ont par exemple prouvé qu'il est possible de retrouver des informations sensibles via l'utilisation du bouton "J'aime" de Facebook (voir ci-après)[151]. Le fait de cliquer sur un bouton "J'aime" de Facebook n'est toutefois pas en lien avec la déduction de données à caractère personnel sensibles comme la consommation de drogues.

b) Le contexte de la collecte des données et la nature de la relation entre la personne concernée et le responsable. La question de savoir si les nouvelles informations que l'on peut découvrir à l'aide d'une analyse de big data (par exemple le risque de fraude) répondent à l'exigence d'utilisation compatible dépendra souvent du contexte légal dans lequel travaille le responsable. Ainsi, l'inspection sociale, un scientifique ou un assureur travaillent toujours dans un contexte réglementaire spécifique qui est différent et qui déterminera la possibilité d'utiliser les nouvelles informations. Si la réglementation est suffisamment prévisible à ce niveau et offre les garanties nécessaires pour la personne concernée, une réutilisation sera possible (par exemple un traitement à des fins scientifiques).

[151] YOUYOU, Wu, KOSINSKI Michal, STILLWELL, D., *Computer-based personality judgments are more accurate than those made by humans*, Department of Psychology, University of Cambridge, 27 janvier 2015. publié sur http://www.pnas.org/content/112/4/1036.full.pdf, cité dans NORTH, A., *How Your Facebook Likes Could Cost You a Job*, 20 janvier 2015. publié sur le blog du New York Times, https://op-talk.blogs.nytimes.com/2015/01/20/how-your-facebook-likes-could-cost-you-a-job/?_r=2.

c) Le fait que des données à caractère personnel sensibles ou non sont traitées (voir ci-après) ;

d) Les éventuelles conséquences négatives (risques) du traitement ultérieur prévu pour les personnes concernées ;

Un exemple possible est la situation où l'intéressé (moyen) est exposé à un risque supérieur à celui prévu initialement dans les finalités de départ. Cela pourrait s'appliquer à une administration fiscale qui promet l'octroi d'une réduction unique contre la communication d'informations supplémentaires sur le patrimoine via une application web (la finalité initiale de cette information est de calculer le montant de la réduction), après quoi le patrimoine est imposé plus efficacement (et généralement davantage) à l'aide d'analyses de big data sur la base des informations obtenues, de sorte que l'utilisateur moyen devra au final payer davantage ;

e) L'existence de garanties et de contrôles adéquats (comme éventuellement le cryptage ou la pseudonymisation), d'autres garanties complémentaires ou de conventions solides contre la réidentification. Si par exemple dans le cadre de l'établissement de modèles prédictifs, un Proof Of Concept est développé dans une première phase (voir la dernière recommandation et le schéma 1 du point G, où le modèle est établi et sa performance est testée sur la base de données les plus anonymes possible et où au moins les identifiants directs sont supprimés, ou que d'autres techniques d'anonymisation ont été utilisées, voir le point C), et que le modèle n'est pas appliqué dans la phase suivante à un autre groupe cible de personnes identifiables (séparation fonctionnelle ou "functional separation"), cela peut constituer un élément pour motiver la compatibilité. Dans le schéma 1, cela signifie que la deuxième phase n'est pas (encore) appliquée.

✓ Recommandation 20

Pour les projets de big data, il est important de décrire au préalable et au maximum toutes les finalités possibles pour le(s) traitement(s) ultérieur(s). En d'autres termes, les domaines de recherche et finalités possibles doivent être délimités le plus possible et de manière préalable.

✓ Recommandation 21

Les responsables qui n'ont pas de finalité opérationnelle directe à l'égard d'individus et souhaitent donc uniquement obtenir des données ou des résultats globaux qui ne sont plus des données à caractère personnel (par exemple données agrégées ou (la performance d') un modèle mathématique) dans un projet de big data peuvent faire traiter leurs données à caractère personnel – à tout le moins pseudonymisées – par un tiers (par exemple un groupe de

recherche académique) qui offre des garanties suffisantes qu'un projet de recherche *scientifique* sera démarré, où les données à caractère personnel seront détruites par le sous-traitant à la fin du projet. Tant l'article 4, § 1, 2° de la LVP[152] que l'article 5 1., b) du RGPD prévoient à cet égard des dispositions particulières quant à la compatibilité de traitements ultérieurs à des fins historiques, statistiques ou scientifiques. Dans la mesure où il est question de recherche scientifique (c'est-à-dire dans la mesure où l'on répond à des conditions pertinentes [153]), aucune donnée à caractère personnel n'est communiquée à des tiers et des garanties suffisantes peuvent être données en matière de sécurité des données à caractère personnel et cela peut constituer une garantie suffisante pour le respect de l'exigence d'utilisation compatible.

Dans la pratique, on observe des tendances qui vont à l'encontre d'une application correcte du principe de finalité, comme :

a) **Une confusion de finalité** (la "lutte contre la fraude" est parfois mentionnée comme finalité générique très vague, où on ne sait pas toujours clairement ce qu'est la "fraude", quels types et gradations de fraude on souhaite combattre sur quelle base (légale)[154])
b) "**Window dressing**" : pratique consistant à ne pas mentionner correctement toutes les finalités ou à présenter une situation autrement au monde extérieur (comme le fait de faire passer une activité de surveillance pour une mesure de sécurité[155])
c) "**Function creep**", où les données collectées pour une finalité déterminée (par exemple facturation de l'énergie et de l'eau à l'aide d'informations dans le registre d'accès de gestionnaires de réseau de distribution[156]) sont utilisées ultérieurement à d'autres fins (par exemple lutte contre la fraude à l'énergie).

---

[152] Voir les articles 2 à 24 inclus de l'arrêté royal du 13 février 2001 *portant exécution de la loi du 8 décembre 1992 relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel*, M.B., 13 mars 2001.

[153] Bien que le considérant 159 du RGPD parte d'une interprétation large de cette notion qui n'exclut pas le financement privé, le RGPD renvoie bien au respect de conditions spécifiques, "*en particulier, en ce qui concerne la publication ou la divulgation d'une autre manière de résultats de la recherche*". Une recherche est scientifique de par la méthode utilisée (observations et mesures objectives et analyse qui utilise la statistique explicative) mais pour pouvoir justifier pleinement ce qualificatif, elle doit aussi avoir une finalité scientifique. Ceci implique : vouloir contribuer à la connaissance scientifique, participer à des publications reconnues par des pairs et ainsi de suite. Voir la page 7 du vade-mecum du chercheur, publié à l'adresse https://www.privacycommission.be/sites/privacycommission/files/documents/vade-mecum-du-chercheur_0.pdf.

[154] D'après l'Ordre des barreaux flamands, la large notion de "fraude fiscale grave, organisée ou non" est décrite de manière vague dans la loi du 11 janvier 1993 et ses éléments constitutifs ne sont pas clairement distinguables de la "fraude fiscale ordinaire", de sorte que l'insécurité juridique règne. Voir http://www.ordeexpress.be/article/44/283/wet-houdende-dringende-bepalingen-inzake-fraudebestrijding.
Voir également la Décision 2004-499 du Conseil constitutionnel français. Cette décision concernait l'intention de reprendre un nouvel article 9 dans la loi vie privée française de 1978 afin de prévoir la possibilité de traiter des données sur des "infractions, condamnations et mesures de sûreté" pour la prévention et la lutte contre la fraude. Ce en faveur de personnes morales victimes d'infractions ou qui interviennent pour le compte de victimes. Le Conseil a considéré ce qui suit à cet égard : "*Considérant que, s'agissant de l'objet et des conditions du mandat en cause, la disposition critiquée n'apporte pas ces précisions ; qu'elle est ambiguë quant aux infractions auxquelles s'applique le terme de « fraude » ; qu'elle laisse indéterminée la question de savoir dans quelle mesure les données traitées pourraient être partagées ou cédées, ou encore si pourraient y figurer des personnes sur lesquelles pèse la simple crainte qu'elles soient capables de commettre une infraction* (…)". Voir http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/les-decisions/acces-par-date/decisions-depuis-1959/2004/2004-499-dc/decision-n-2004-499-dc-du-29-juillet-2004.904.html.

[155] Voir la discussion sur la finalité réelle du cookie datr de Facebook. D'après la Commission, il s'agit d'une surveillance des utilisateurs de sites web ; d'après Facebook, il s'agit de mesures de sécurité. Voir Facebook wint veldslag, maar privacystrijd is nog lang niet gestreden, de redactie, 29 juin 2016, http://deredactie.be/cm/vrtnieuws/cultuur%2Ben%2Bmedia/media/2.37414.

[156] Les gestionnaires de réseau de distribution (EANDIS, INFRAX, …) enregistrent les données de raccordements au réseau de gaz et d'électricité (notamment) de clients résidentiels dans ce qu'on appelle le registre d'accès. Dans ce registre, ils inscrivent les données techniques de chaque raccordement ainsi que les données administratives du client et son fournisseur d'énergie.

Une application spécifique réside dans la réutilisation d' "open data" pour des projets de "big data". D'une part, il est essentiel pour la science que des données soient disponibles et puissent être partagées[157]. D'autre part, le fait que des autorités publiques et le secteur privé rendent publiques des données peut aussi impliquer de nouveaux risques pour les droits et libertés des personnes concernées. Une réutilisation d'open data pour des projets de big data peut dans ce cas poser problème à l'égard des prévisions raisonnables des intéressés.

I. Principe de légalité

Un bon cadre législatif pour les services publics qui utilisent les analyses de big data ou les "patrouilles en ligne" est important. Il semble toutefois exister une "distorsion" dans les cadres législatifs. Dans la pratique, l'attention est surtout centrée sur la collecte et le partage de données (par exemple l'accent est mis sur la constitution d'un data warehouse (entrepôt de données)[158]). Trop souvent encore, un cadre général pour l'analyse et l'application des résultats aux personnes [159] par les services d'inspection fait défaut (garanties pour des analyses de big data de qualité et prévention du déterminisme des données, force probante des constatations, droits des personnes concernées à cet égard, ...).

✓ **Recommandation 22**

Le législateur devrait consacrer davantage d'attention à l'obligation de réaliser des analyses de big data de haute qualité et à l'application d'analyses de big data à des personnes physiques par des services publics ou d'autres services qui remplissent une mission d'intérêt général. Cela est possible en tenant compte des éléments figurant dans le présent rapport. Dans ce cadre, on peut en particulier faire référence à ce qui a été précisé au point D.1.

J. Légitimité / Possibilités de considérer les analyses de big data comme un traitement légitime

Les analyses de big data ne sont pas des traitements qui seraient a priori et de par leur nature (il)légaux mais ce ne sont des traitements légitimes de données à caractère personnel que dans la mesure où un but légitime est poursuivi et où un juste équilibre est trouvé entre la protection de la vie privée et des données à caractère personnel d'une part et l'impact sur la société d'autre part.

La LVP prévoit actuellement quelques possibilités à l'article 5. Cet aspect est abordé à l'article 6 du RGPD. Toutefois, dans le cadre de certaines possibilités, dans certaines circonstances, des problèmes peuvent survenir pour offrir aux

---

[157] de Montjoye Y. A., *Unique in the shopping mall:* On the re-identifiability of credit card metadata, Science vol 347, 30 janvier 2015 ; http://science.sciencemag.org/content/347/6221/536; 536-537.
[158] Voir par exemple la loi du 3 août 2012 *portant dispositions relatives aux traitements de données à caractère personnel réalisés par le Service public fédéral Finances dans le cadre de ses missions*, M.B., 24 août 2012.
[159] BROEDERS, D. et SCHRIJVERS, E., *Big Data in een vrije en veilige samenleving*, 20 mai 2016, http://njb.nl/njv-jaarvergaderingen/jaarvergadering-2016/artikelen/big-data-in-een-vrije-en-veilige-samenleving.19799.lynkx.

traitements de big data une légitimité suffisante (afin qu'il s'agisse de traitements légitimes).

Pour le consentement de la personne concernée, la question est en effet de savoir si celle-ci peut donner son consentement en toute connaissance de cause[160] pour un traitement dont l'impact réel peut difficilement être connu (voir également au point L quelles informations clés doivent être communiquées de manière transparente) même si tous les efforts sont réalisés afin d'expliquer les différents aspects du traitement dans un langage aussi compréhensible que possible. Dans le cadre de projets de big data, la question est de savoir si cela est toujours tout à fait possible et s'il est même toujours possible de prévoir à l'avance tous les effets (secondaires).

La protection de la vie privée et des données à caractère personnel n'est pas absolue et exige souvent une pondération des intérêts par rapport à d'autres intérêts légitimes d'intérêt plutôt général (par exemple une étude scientifique ou la liberté de commerce et d'industrie) ou des intérêts légitimes d'un responsable ou d'un tiers. Ces intérêts peuvent offrir un fondement juridique au traitement, à condition qu'il y ait un équilibre des intérêts avec la personne concernée, en tenant compte des prévisions raisonnables de cette dernière. Il faut aussi examiner s'il n'y a pas d'incidence négative sur certains groupes de population (voir ci-après au point E.2)[161].

Motiver la légitimité des projets de big data à l'aide d'un intérêt légitime d'un responsable ou via le consentement des clients n'est en pratique pas évident, eu égard à la complexité, au pouvoir prédictif potentiel et à l'impact de certains modèles de big data prescriptifs sur la personne concernée. Il est également essentiel de prévoir des garanties complémentaires qui estiment ou, le cas échéant, traitent l'impact social d'analyses de big data (voir également le point E.2. ci-avant). La réalisation d'une analyse d'impact relative à la protection des données constituera une amorce dans ce cadre, dans la mesure où lors de cette analyse, on ne se limite pas à analyser le nombre de cas de problèmes rapportés que les individus ont rencontrés à la suite d'analyses de big data.

✓ Recommandation 23

Si le responsable invoque le consentement, celui-ci doit pouvoir être retiré facilement et il ne vaudra pas dans le cadre d'un déséquilibre des forces[162] clair entre la personne concernée et le responsable ou le sous-traitant, par exemple parce que le responsable fournit le service dominant (ou le seul service) sur le

[160] Considérant 42 du RGPD et la FAQ relative au consentement sur le site Internet de la Commission (https://www.privacycommission.be/fr/faq-page/10027#t10027n20395).
[161] G.H. Evers, *In de schaduw van de rechtsstaat : profilering en nudging door de overheid*, Computerrecht, 2016/98, juin 2016, p. 169.
[162] D'après le Considérant 42 du RGPD, le consentement ne peut pas être considéré comme étant donné librement si la personne concernée n'a pas de choix véritable ou libre ou qu'elle ne peut pas refuser son consentement ou le retirer sans conséquences négatives. Voir également le Considérant 43 du RGPD qui dispose que le consentement n'offre pas un fondement juridique valable s'il "*existe un déséquilibre manifeste entre la personne concernée et le responsable du traitement*". C'est ce qui découle de l'exigence de liberté du consentement à l'article 4, 11) du RGPD. Voir également les pages 12-16 de l'avis n° 15/2011 du Groupe 29 du 13 juillet 2011 sur la définition de consentement http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf et la FAQ sur le site Internet de la Commission (https://www.privacycommission.be/fr/faq-page/10027#t10027n20395).

marché. Le responsable devra prouver qu'il n'y a pas de déséquilibre des forces ou que ce déséquilibre ne peut pas influencer le consentement de la personne concernée.

K. Obligation d'information

Le responsable a en principe une obligation d'information en vertu de l'article 9 de la LVP et des articles 13 et 14 du RGPD[163]. Ainsi, la personne concernée doit être informée de points spécifiques et bien définis dont l'existence d'un profilage et des conséquences de celui-ci[164].

Des exceptions à l'obligation d'information sont uniquement prévues dans un nombre limité de cas par la législation (par exemple la législation sur le blanchiment, la lutte contre la fraude à la sécurité sociale, le traitement à des fins policières, une enquête fiscale, …).

La distinction entre le droit à la protection des données et le droit à la vie privée (notamment les articles 8 de la CEDH et 22 de la Constitution) est importante car, même si l'obligation d'information ou le droit d'accès ne s'applique pas en vertu des exceptions susmentionnées, une obligation générale de transparence est quand même d'application en vertu du droit à la vie privée via ce qu'on appelle l'exigence de "prévisibilité" de la législation[165] dans le cadre d'ingérences dans la vie privée. La notion de législation doit ici être comprise au sens large (voir le point L ci-dessous).

L. Transparence

La transparence est une obligation légale qui découle du droit à la vie privée. La transparence signifie que des informations et des communications liées au traitement doivent être aisément accessibles et compréhensibles/claires. Ce qu'on appelle l'exigence de "prévisibilité" de la législation constitue également une exigence classique en matière de vie privée[166], qui implique une obligation de transparence.

Les analyses de big data et l'utilisation d'algorithmes exigent donc, en règle générale, de la transparence, même si le responsable est dispensé de l'obligation d'information (voir ci-avant).

✓ Recommandation 24

Dans la pratique, il faut faire une distinction entre le respect (l'obligation de respecter) ou non de **l'obligation d'information d'une part** et **l'obligation de transparence**[167] **d'autre part**. La transparence doit concerner toutes les

[163] Articles 13 et 14 du RGPD.
[164] Considérant 60 du RGPD.
[165] Article 8.2. de la CEDH, tel que repris par l'article 22 de la Constitution et l'article 7 de la Charte EU.
[166] Article 8.2. de la CEDH, tel que repris par l'article 22 de la Constitution et l'article 7 de la Charte EU.
[167] Parfois, on consacre uniquement de l'attention à l'obligation d'information et aux exceptions à cette obligation en vertu de la législation sur la protection des données et on oublie l'obligation de transparence en vertu de la législation en matière de vie privée. Si par exemple des banques ou des organismes d'assurance sont dispensés de l'obligation

phases du traitement de données (voir la Partie 1) et plus spécifiquement aussi bien la phase d'apprentissage et la phase de test que la phase de l'application des modèles (voir le schéma 1) dans le cadre de l'établissement de modèles prédictifs.

Cependant, la plupart des analyses de big data et des applications d'algorithmes dans le secteur privé et le secteur public ne sont actuellement pas souvent basées sur un cadre légal clair, sur des clauses contractuelles suffisamment claires ou sur une politique externe de confidentialité. Les conditions générales d'institutions financières renvoient souvent uniquement à des formulations et des termes très généraux (la possibilité d'établir des "profils" et de "réaliser des analyses de données"), alors que la nature exacte et la puissance prédictive des applications ne sont souvent pas claires du tout.

Les caractéristiques des individus deviennent donc toujours plus transparentes pour les responsables, à l'aide d'analyses de big data, alors que des algorithmes sont souvent caractérisés par une absence de transparence ou une opacité ("opacity") à l'égard des personnes concernées et des contrôleurs[168].

L'opacité du code source ou du fonctionnement détaillé ou de l'utilisation spécifique d'algorithmes peut parfois être souhaitable et même nécessaire. Cette opacité peut en effet garantir l'efficacité d'enquêtes sur la fraude dans l'intérêt général, servir les droits ou libertés de certaines parties (par exemple des algorithmes de cryptage), y compris le secret d'entreprise ou la propriété intellectuelle et le droit d'auteur qui protège le logiciel[169]. Dans ce cas, on peut choisir de ne partager certaines informations qu'avec le contrôleur.
L'obligation de transparence n'est donc pas absolue mais doit s'appliquer au cas par cas pour rechercher un bon équilibre qui tient compte d'autres intérêts.
Dès lors, la Commission ne plaide pas pour une transparence absolue (par exemple au moyen de la publication systématique du code source ou de la description d'algorithmes de manière à ce qu'ils puissent être tout à fait reproduits) mais pour un bon équilibre entre le secret (qui doit alors servir un intérêt légitime ou social) d'une part et la transparence d'autre part. La communication d'autant d' "informations clés" que possible (voir ci-après) peut contribuer à atteindre cet équilibre.

Une opacité inappropriée ou un manque de transparence peut toutefois également faire l'objet d'abus pour contourner la législation en matière de vie privée, de protection des données, de protection des consommateurs ou de compétition en réduisant la protection légale ou les droits des personnes concernées ou en les rendant insignifiants (voir ci-après aussi les points N, O et P relatifs aux droits des personnes concernées). Si des clients ou des citoyens n'ont qu'une explication limitée au sujet de la (qualité de la) logique à l'initiative de certaines décisions prises les concernant, il est également très difficile pour eux de reconnaître des décisions sous-optimales (ou erronées), ne leur permettant pas dans les faits de faire valoir leurs droits (par exemple le droit de

---

d'information en vertu de la législation sur le blanchiment, cela ne signifie pas que la législation relative aux processus de données peut être tout à fait non transparente.
[168] BURRELL, Jenna, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, janvier 2016, HTTP://bds.sagepub.com/content/3/1/2053951715622512 ; page 8 (challenge 2) de Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, mai 2016, publié sur https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
[169] Considérant 63 du RGPD.

rectification).

Pour maintenir une possibilité de contrôle par les personnes concernées de leurs données à caractère personnel[170], il est très important que les responsables qui utilisent des analyses de big data maintiennent un niveau minimal de transparence, surtout s'il s'agit de modèles (prédictifs ou prescriptifs) qui peuvent avoir un impact considérable sur les droits et libertés des personnes concernées. Cela doit leur permettre de se (faire) défendre le cas échéant contre des décisions qui ont été prises les concernant. On aborde ci-après quelques éléments devant être mis à disposition en priorité et en toute transparence.

L.1. Transparence d' "informations clés" dans le cadre d'analyses de big data

✓ **Recommandation 25**

Si des analyses de big data constituent une ingérence dans la vie privée (par exemple parce que des prévisions sont réalisées concernant des personnes ou qu'on essaie d'influencer un comportement), ces ingérences doivent être suffisamment prévisibles. Après pondération d'autres intérêts sociaux, la Commission estime que cette prévisibilité peut être apportée par la transparence des éléments de base ou les informations clés[171] définis ci-dessous à l'égard des personnes concernées.

Comme déjà expliqué ci-avant, cela ne signifie donc pas par exemple que des modèles (par exemple des modèles de score de crédit) doivent être publiés dans tous les cas et la Commission estime que cela n'est pas incompatible avec les droits ou libertés des autres, incluant le secret d'entreprise ou la propriété intellectuelle et notamment le droit d'auteur qui protège le logiciel[172]. Dans la législation relative à l'environnement et à l'énergie, la publication d'informations essentielles sur un produit ou service (par exemple un label énergétique d'un réfrigérateur) par les fabricants n'est pas non plus expliquée comme étant une violation de leur droit de propriété intellectuelle, mais comme une opportunité d'innovation et de diversification de leurs produits.

a) Finalité et responsabilité

Le fait d'effectuer des analyses de big data et des choix liés à un modèle requiert toujours une implication humaine. Pour l'application de la réglementation de protection de la vie privée et des données, il est également essentiel que les finalités du projet soient déterminées et connues. Les points de

[170] Voir la page 8 de l'avis du CEPD du 19 novembre 2015, *Relever les défis des données massives, Un appel à la transparence, au contrôle par l'utilisateur, à la protection des données dès la conception et à la reddition de comptes*, publié sur https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_FR.pdf

[171] DE BOT, D. et DE HERT, P., article *22 Grondwet en het onderscheid tussen privacyrecht en gegevensbeschermingsrecht. Een formele wet is niet altijd nodig wanneer de overheid persoonsgegevens verwerkt, maar toch vaak*, CDPK, 2013, 358.

contact que l'on peut interpeller, qui sont (co)responsables à cet effet et qui peuvent prendre (adapter) des décisions (par exemple le délégué à la protection des données ("DPO" en anglais)[173]), doivent être rendus publics. Sans cette transparence, il ne peut y avoir non plus aucune obligation de responsabilité ("accountability") et la responsabilité des analyses de big data ne peut pas être attribuée.

b) Les données utilisées

Une définition des données (à caractère personnel) qui sont utilisées pour l'apprentissage, le test et l'utilisation opérationnelle d'un modèle (voir également le schéma 1) doit être considérée comme une information clé[174]. Dans ce cadre, il faut restituer de manière aussi reproductible que possible la façon dont les données ont été obtenues et quelles en sont les caractéristiques. Il ne s'agit pas ici uniquement des sources de données choisies et de la manière dont les données ont été collectées (via des enquêtes, des applications en ligne ou mobiles, des données ouvertes (open data), des médias sociaux, des couplages entre plusieurs sources de données, ...) mais aussi des variables collectées et de la manière dont celles-ci ont été mesurées, de la définition et de la détermination des étiquettes de classification, de l'éventuel pré-traitement, de la manière dont les points dans l'ensemble d'apprentissage et de test ont été "extraits" de la population totale dont on souhaite réaliser des prévisions, de la qualité des données, des incertitudes, de l'ancienneté des données, de la fréquence de réapprentissage, ... Dans ce cadre, il faut également reproduire la manière dont on a pris les précautions nécessaires afin que les données d'apprentissage et de test soient (et restent) représentatives du groupe cible pour lequel on entend réaliser des prévisions ou prendre des décisions.

c) Le Modèle

La performance ou la précision du modèle[175] (voir le schéma 1) telle que testée sur des données de test indépendantes et représentatives doit être tenue à la disposition des personnes concernées[176].

Il est également crucial que la transparence soit faite concernant les principaux choix méthodologiques, comme les choix en matière d'algorithmes (d'apprentissage) et de structure du modèle, le mode de détermination d'éventuels (hyper)paramètres (tels que par exemple le cut-off du modèle[177]), les variables ou caractéristiques qui sont prises en considération pour être incluses dans le modèle et la manière dont les variables/caractéristiques ont été sélectionnées dans le modèle définitif, …

---

[173] Articles 37 et suivants du RGPD.

[174] Voir les points 29 à 33 inclus de l'avis n° 45/2013 du 2 octobre 2013 concernant le Code wallon de l'Agriculture ainsi que le point 15 de l'avis n° 08/2005 du 25 mai 2005 concernant un avant-projet de loi relatif à l'analyse de la menace.

[175] Par exemple le nombre de true positives, de true negatives, de false positives et de false negatives (permettant de déterminer d'autres mesures de performance comme l'exactitude, la sensibilité, la spécificité, …), l'aire sous la courbe ROC, …

[176] Dans la littérature, on fait la comparaison avec la publication des résultats de crash tests de voitures, ce qui ne signifie pas nécessairement que l'on doive rendre public le mode de construction de la voiture. Voir la p. 59 de DIAKOPOULOS, N., *Accountability in Algorithmic Decision Making*, février 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/32/2016).Responsabilité-O.algorithmic-decision-making-Final.pdf.

[177] Il s'agit du seuil que l'on utilise pour le produit de sortie (continu) du modèle afin de ranger une personne dans une des deux classes.

Il est souhaitable que la description des algorithmes décisionnels sous-jacents proprement dits soit aussi mise à disposition si cela est possible dans la pratique (c'est donc le modèle final qui est utilisé après l'application de la méthodologie choisie dans la Phase 2 du schéma 1 - il s'agit donc ici de la logique avec la relation concrète input-output).

Cette information doit rendre les traitements de données pour des tiers (contrôleurs, personnes concernées, ...) aussi reproductibles que possible et leur permettre d'éventuellement se (faire) défendre contre des décisions méthodologiques sous-optimales ou des modèles ou des analyses peu fiables. Cela ne signifie pas nécessairement que l'on doive toujours rendre complètement reproductibles (par exemple via le code source) la méthodologie et les modèles utilisés et les rendre publics dans tous les détails. Dans ce cadre, comme cela a déjà été précisé, il faut réaliser une pondération avec d'autres intérêts légitimes tels que les intérêts commerciaux et de propriété intellectuelle, la protection des secrets d'entreprise, le savoir-faire et la sécurisation du traitement. Des informations sensibles ou critiques relatives à l'infrastructure informatique ou au réseau de l'entreprise par exemple peuvent, en raison de risques de sécurité, faire l'objet d'une exception à la transparence publique. Sur demande, le responsable doit toutefois toujours offrir une complète transparence vis-à-vis des contrôleurs (avec tous les détails afin que le traitement puisse être tout à fait reconstruit = reproductibilité complète) quant à la manière dont certains résultats ont été obtenus.

Les éléments décrits ci-dessus, et plus particulièrement la relation concrète input-output qui est à l'initiative d'une décision ou d'une prévision, doivent également (voir le point J ci-dessus) être expliqués aux personnes concernées dans un langage aussi compréhensible que possible. On peut éventuellement opter ici pour une approche par couches sur un site Internet où sont d'abord proposés les éléments clés (courte information qui est fortement résumée), après quoi, il est possible de cliquer sur un lien si on est intéressé par des informations plus approfondies.

d) Les conséquences

Il faut toujours indiquer clairement si une décision a été basée (notamment) sur des analyses de big data (par exemple si des modèles prédictifs ont été utilisés).

Les conséquences à prévoir pour les (différentes catégories de) personnes concernées lors de l'utilisation ou d'une décision déterminée du modèle doivent être transparentes (par exemple augmentation ou réduction d'une prime d'assurance[178], déduction ou prévision de certaines informations, ...).

Le résultat de la pondération continue des risques (voir ci-dessus le point B concernant le RGPD) doit être publié, ce qui n'implique pas que toutes les informations dans ce cadre doivent être rendues publiques.

Pour les "open data", il importe que le responsable, avant que les "open data" soient mises à disposition, soit également transparent quant à son examen visant à connaître les risques de réidentification et permettant d'avoir une idée

[178] Simon, F; insurance exec: Big data allows better understanding of risk, 9 juin 2016, http://eurac.tv/25GO.

en externe de l'unicité des ensembles de données et de la possibilité pour des tiers (par exemple des sociétés d'informations commerciales ou des data brokers (courtiers de données)) d'isoler des données jusqu'à un niveau individuel (voir également le point C).

M. Protection de données à caractère personnel sensibles

Établir des "corrélations" avec ou sur la base de données à caractère personnel sensibles (comme des données relatives à la race, à l'origine ethnique, aux opinions politiques, à la religion ou aux convictions philosophiques, à l'appartenance syndicale, au statut génétique ou à l'état de santé, ou à l'orientation sexuelle [179] ou relatives aux condamnations pénales et aux infractions ou aux mesures de sûreté connexes[180]) est en principe interdit, à moins qu'une exception légale existe et que des garanties soient prévues[181]. Vu le caractère complexe, souvent peu transparent des analyses de big data (voir ci-avant le point J), le "consentement" libre, spécifique, explicite, reposant sur une information préalable de la personne concernée ne sera pas toujours disponible en tant que motif d'exception[182] pour légitimer des corrélations avec ou sur la base de données à caractère personnel sensibles.

Les analyses de big data peuvent également avoir un **impact transformateur** sur des données à caractère personnel. Cela signifie que des données à caractère personnel en principe non sensibles ou des données apparemment inoffensives peuvent potentiellement être converties en (une corrélation de) données à caractère personnel sensibles. Au départ des "likes" de Facebook publiquement disponibles, on pourrait par exemple établir un modèle informatique qui peut décrire la personnalité d'un individu de manière plus précise que cela n'est possible pour les amis, les collègues, les conjoints ou la famille[183]. L'orientation sexuelle, l'origine ethnique, les convictions religieuses et politiques et la consommation de substances addictives telles que l'alcool, le tabac et les drogues peuvent aussi être déduits automatiquement de "likes" de Facebook [184] . Les analyses de big data (d'une combinaison) de données quotidiennes comme l'utilisation de certains produits de soins, la taille des vêtements achetés, les données de localisation, les transactions financières, ... peuvent souvent permettre d'établir une corrélation avec par exemple des données de santé (par exemple la présence d'une maladie ou d'une grossesse, le risque de décès, ...) ou un phénomène de criminalité.

---

[179] Considérants 51 et 71 du RGPD.
[180] En vertu de l'article 10 du RGPD, cette définition plus stricte est utilisée par rapport à la définition bien plus large de l'article 8 de la LVP.
[181] Voir l'article 9, alinéa 2 du RGPD ; on pet citer comme exemple une étude des universités d'Anvers et d'Amsterdam sur le comportement Twitter au sein d'un cercle social déterminé, mis en corrélation avec le comportement électoral. VAN GYSEL, C., GOETHALS, B. et DE RIJKE, M., *Determining the Presence of Political Parties in Social Circles*, disponible sur https://staff.fnwi.uva.nl/m.derijke/content/publications/icwsm2015-sp-christophe.pdf. La situation est différente s'il s'agit d'un bureau d'étude de marché privé qui effectuerait de telles études, sans qu'une exception légale ne soit prévue.
[182] L'article 8 de la LVP ne prévoit pas le consentement écrit de la personne concernée comme motif d'exception, contrairement aux articles 6 et 7 de la LVP, et l'article 9.2. du RGPD renforce l'exigence de consentement en un consentement "explicite".
[183] YOUYOU, Wu, KOSINSKI Michal, STILLWELL, D., *Computer-based personality judgments are more accurate than those made by humans*, Department of Psychology, University of Cambridge, 27 janvier 2015. publié sur http://www.pnas.org/content/112/4/1036.full.pdf, cité dans NORTH, A., *How Your Facebook Likes Could Cost You a Job*, 20 janvier 2015, publié sur le blog du New York Times, https://op-talk.blogs.nytimes.com/2015/01/20/how-your-facebook-likes-could-cost-you-a-job/?_r=2.
[184] KOSINSKI Michal, STILLWELL, D. et GRAEPEL, T., *Private Traits and attributes are predictable from digital records of human behavior*, Free School lane, The Psychometrics Centre, University of Cambridge, 9 avril 2013, publié sur http://www.pnas.org/content/110/15/5802.full.pdf.

Ce qui est précisément fait ou pas sous le dénominateur d' "analyses de big data", "data mining", "profilage", "traitement à des fins statistiques", ... avec les traces numériques quotidiennes (par exemple l'analyse des traces numériques par une application de paiement, l'analyse de l'achat de biens par un supermarché, le comportement de navigation et d'achat en ligne, ...) devient de plus en plus important si nous souhaitons protéger certains traits de personnalité sensibles (par exemple le profil psychologique ou l'orientation sexuelle). À nouveau, la transparence sera cruciale.

✓ **Recommandation 26**

L'interdiction de traitement de données à caractère personnel sensibles implique également une interdiction de contourner la protection des données à caractère personnel sensibles en utilisant des analyses de big data (transformatrices) à l'égard de données à caractère personnel non sensibles, surtout si la personne concernée a choisi de ne pas communiquer ces caractéristiques.

✓ **Recommandation 27**

Dans le secteur public, le législateur devrait prévoir une surveillance accrue dès qu'il est question de corrélations avec ou sur la base de données à caractère personnel sensibles (par exemple prévention ou détection de la criminalité, de la fraude). Cette surveillance accrue peut prendre la forme d'une autorisation préalable par un contrôleur compétent (par exemple l'Organe de contrôle de la gestion de l'information policière (COC)).

N. Limitations et mesures adéquates en cas de décisions automatisées

La technologie de l'information joue un rôle de plus en plus grand dans les processus décisionnels et prend de plus en plus la place de l'homme[185]. L'informatisation croissante des processus décisionnels a fait apparaître une tendance à réduire l'importance et le rôle de l'intervention réelle et consciente d'individus dans les processus décisionnels. Lorsque de telles décisions automatisées ont un impact manifeste sur des personnes, le RGPD et la jurisprudence européenne fixent des limites[186].

Les articles 12*bis* de la LVP et 22.1. du RGPD disposent que "*la personne*

[185] Un exemple fictif mais réaliste concerne l'utilisation d'un logiciel par un professeur d'université afin d'évaluer le risque de plagiat dans un travail d'un étudiant. Dans la mesure où on ne tient pas suffisamment compte de la marge d'erreur inhérente d'une utilisation exagérée de ce logiciel et de la possibilité de l'étudiant d'être entendu dans le cas d'une décision de plagiat, cela peut parfois conduire, dans la pratique, à une décision automatisée ayant un impact considérable sur la personne concernée.

[186] Cour de justice, Avis 1/2015 du 26 juillet 2017. Aux paragraphes 131, 168 et suivants de cet avis, on se réfère aux modèles et critères préétablis avec une "marge d'erreur significative" et au choix des banques de données pour la comparaison de fichiers des analyses PNR automatisées, qui déterminent le degré d'ingérence dans la vie privée de la personne concernée. La Cour souhaitait que la fiabilité et l'actualité de ces modèles, critères et banques de données préétablis fassent partie du contrôle "joint review" suivant en vertu de l'Accord PNR, afin de vérifier la nécessité et caractère non discriminatoire (§ 174).

concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire".

Il n'y a pas d'interdiction absolue de procéder à des décisions automatisées, étant donné que l'article 22.2. du RGPD prévoit des exceptions, comme le consentement explicite de la personne concernée, une nécessité pour la conclusion ou l'exécution d'un contrat avec la personne concernée ou une décision basée sur le profil qui repose sur une obligation en vertu du droit de l'UE ou d'un État membre avec des garanties appropriées. On peut citer comme exemple la notification d'une banque à la cellule anti-blanchiment dans le cadre de laquelle le risque de la transaction a été défini sur la base de l'enquête client (ce qu'on appelle la "customer due diligence"), en vertu d'obligations reprises aux articles 7 et suivants de la loi du 11 janvier 1993 *relative à la prévention de l'utilisation du système financier aux fins du blanchiment de capitaux et du financement du terrorisme*[187].

Une intervention visible d'une personne physique dans le processus décisionnel n'est pas toujours suffisante. Une évaluation autonome et humaine manifeste dans le cadre de l'utilisation de technologies ICT dans la procédure décisionnelle constitue une exigence minimale, ce qui découle du considérant 71 et de l'article 22.3. du RGPD.

✓ Recommandation 28

Des dérogations manifestes prises par des personnes physiques doivent être disponibles dans l'application systématique du résultat d'analyses de big data qui ont un impact considérable pour les droits individuels des personnes concernées.

Une prise de décision automatisée et un profilage basé sur des catégories particulières de données à caractère personnel ("données à caractère personnel sensibles") peuvent, selon le RGPD, exclusivement être autorisés si des conditions spécifiques sont remplies[188], comme la présence d'un consentement explicite, la communication d'informations spécifiques à la personne concernée[189] et/ou l'octroi d'un droit d'accès (adapté) (voir également le point O). Concrètement, il faudrait définir ce que les droits d'accès, de rectification et d'opposition impliquent dans le cadre d'analyses de big data : la personne concernée a-t-elle ou non accès à l'algorithme ou aux éléments essentiels des analyses de big data (voir le point L.1. ci-dessus) ? Si ces explications concrètes font défaut, ces droits risquent de rester lettre morte. On peut y parvenir par exemple en insistant plus fortement sur l'obligation d'information, par analogie avec les règles "e-privacy" des articles 122 et 123 de la loi du 13 juin 2005

[187] Loi du 11 janvier 1993 *relative à la prévention de l'utilisation du système financier aux fins du blanchiment de capitaux et du financement du terrorisme*, telle que modifiée pour la dernière fois par la loi du 13 mars 2016, M.B. du 9 février 1993.
[188] Considérant 71 du RGPD.
[189] Voir le considérant 71 du RGPD, l'article 13.2. f) du RGPD et l'article 14.2. g) du RGPD concernant le devoir d'information.

relative aux communications électroniques en ce qui concerne les données de trafic et les données de géolocalisation.

O. Droits d'accès et de rectification, droit d'effacement des données

La personne concernée a un droit d'accès aux données à caractère personnel la concernant [190] , et le droit d'obtenir du responsable du traitement la rectification [191] des données à caractère personnel la concernant qui sont inexactes. En vertu du RGPD, le responsable du traitement doit accéder à cette requête "dans les meilleurs délais", et en tout cas dans le mois de sa réception. Il est possible de prolonger de deux mois ce délai pour les requêtes complexes et en fonction du nombre de requêtes que le responsable reçoit [192] . Le responsable du traitement informe la personne concernée d'une telle prolongation dans le mois qui suit la réception de la requête.

L'article 15.1. h) du RGPD dispose que l'accès ne concerne pas uniquement l'existence de décisions automatisées, mais également des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée. Voir aussi dans ce cadre les recommandations dans le présent rapport au point L.1. (Transparence des informations clés).

Selon la Commission, les termes "au moins en pareils cas"[193] de l'article 15.1. h) du RGPD ne peuvent pas être interprétés de manière à ce qu'il ne s'agisse que du traitement de données à caractère personnel sensibles et des cas impliquant un impact manifeste sur les personnes concernées. Dans le cas où il est par exemple question d'un risque de traitement déloyal ou d'une discrimination ou dans le cas où la non communication de ces informations ferait des informations bel et bien fournies des informations trompeuses ou vides de sens, par exemple quand le responsable communique uniquement que le traitement a lieu "à des fins statistiques", sans expliquer l'impact des décisions), les informations figurant à l'article 15.1. h) du RGPD doivent, selon la Commission, toujours être fournies.

L'article 17 du RGPD ajoute un droit à l'effacement ("droit à l'oubli"). À cet égard, il est pertinent de préciser que certaines structures de big data qui transfèrent des données de manière séquentielle (comme méthode pour stocker efficacement des données) peuvent entraîner un allongement de la durée nécessaire à l'effacement physique des données[194]. Dans le cadre du droit à l'effacement, cela peut constituer un problème lorsque l'effacement ne peut pas être assuré "dans les meilleurs délais"[195].

---

[190] Article 10 de la LVP et article 15 du RGPD.
[191] Article 12 de la LVP et article 16 du RGPD.
[192] Article 12.3 du RGPD.
[193] Dans les cas d'une décision admise par le RGPD, fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire. Il en va de même pour les catégories particulières de données à caractère personnel pour lesquelles des garanties suffisantes ont été prévues (les cas visés aux articles 22.1. et 22.4. du RGPD).
[194] Voir https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (voir le point "space reclamation").
[195] Article 17.1 du RGPD.

P. Droit d'opposition

L'article 21.2. du RGPD dispose que "*lorsque les données à caractère personnel sont traitées à des fins de prospection, la personne concernée a le droit de s'opposer à tout moment au traitement des données à caractère personnel la concernant à de telles fins de prospection, y compris au profilage dans la mesure où il est lié à une telle prospection*".

On peut se demander ce que le droit d'opposition signifie concrètement dans le cadre d'analyses de big data. L'article 21 du RGPD n'ouvre aucun droit général de ne pas être soumis au profilage ou à des analyses de big data (voir également au point N). La question de savoir s'il existe un droit d'opposition dépendra du type de traitement qui est visé et il faut examiner quels intérêts légitimes prévalent (par exemple marketing direct contre recherche scientifique ; profilage par des services publics avec une base légale ou réglementaire contre profilage par une société commerciale).

Il n'empêche de toute façon que certaines décisions basées sur des résultats d'analyses de big data ou de comparaisons de fichiers pourront être perçues par les personnes concernées comme incorrectes, injustes ou discriminatoires (par exemple le refus d'un accès à un festival sur la base d'une comparaison de fichiers).

✓ **Recommandation 29**

En raison de la difficulté d'application du droit d'opposition, il semble parfois qu'il soit recommandé à la personne concernée de recourir à la possibilité d'une procédure devant le juge des référés. Il est toutefois aussi important de prévoir une possibilité de recours auprès d'une autorité compétente contre des décisions qui sont (en partie) basées sur des analyses de big data, dont un recours auprès des autorités ou organes de contrôle compétents pour surveiller la protection de clients et de consommateurs dans des secteurs particuliers (par exemple le secteur des télécommunications et le secteur financier), ou une législation particulière (par exemple le respect des dispositions anti-discrimination). Dans ce cadre, on peut également penser à une représentation des personnes concernées (article 80 du RGPD) ou (pour un groupe de personnes concernées qui sont consommatrices et ont subi un dommage) à la possibilité d'une action en réparation collective[196].

Q. Responsabilité du traitement

Dans le cadre d'analyses de big data, il est de plus en plus question d'un grand nombre d'acteurs au niveau national et/ou international, ce qui rend l'attribution des responsabilités plus complexe, selon que plusieurs parties assurent différentes tâches dans le secteur privé et/ou public[197], comme la collecte et la

[196] Sur la base de l'article XVII.36 et de l'article 37 du Code de droit économique ("CDE").
[197] VAN DER SLOOT, B. et VAN SCHENDEL, S. Wetenschappelijke Raad voor het Regeringsbeleid, *International and Comparative Legal study on Big Data*, p. 43, publié à l'adresse http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf.

fourniture de données[198], la fourniture de logiciels et/ou de matériel, ... Afin de déterminer si une partie est responsable du traitement, la Commission peut à cet égard tenir compte de tous les facteurs pertinents, mentionnés dans la définition du responsable du traitement (article 4, 7 du RGPD).

✓ **Recommandation 30**

Dans la réglementation et les dispositions contractuelles, le(s) responsable(s) des traitements de big data doi(ven)t toujours être mentionné(s) clairement.

R. Contrôle des traitements

Une interception à grande échelle (dans le cadre du contrôle ou de la surveillance de personnes) de données qui peuvent concerner tous les utilisateurs potentiels d'un service constitue une ingérence dans la vie privée et exige toujours un contrôle indépendant[199] pour surveiller la nécessité. On peut ici faire référence à des projets et applications dans le secteur public ("predictive policing"), dans le secteur privé ("deep packet inspection" par les opérateurs de télécommunications) [200] et à la privatisation élargie de la prévention de la fraude[201] et de la prévention de la criminalité.

Un risque élevé pour les droits et libertés des personnes physiques peut par exemple aussi résider dans l'utilisation d'algorithmes dans la surveillance par les employeurs ou d'algorithmes dans une application financière de smartphone pour les mineurs[202].

Outre une autorisation préalable par un organe de contrôle indépendant [203] (dans le secteur public), les garanties suivantes peuvent être citées :

- prévoir des audits par des organisations indépendantes à l'égard des systèmes décisionnels qui sont (en partie) basés sur des algorithmes ;
- prévoir des systèmes de modélisation interactive (feed-back par les personnes concernées dont des données sont traitées par des algorithmes et qui permet d'adapter l'algorithme - voir également le point E.2.) ;
- prévoir une organisation indépendante ("Trusted Third Party" - TTP, ou tiers de confiance) qui peut exécuter/garantir une analyse neutre des données selon les règles de l'art (par exemple par une recherche

---

[198] Voir l'article suivant sur le rôle des data brokers dans le fonctionnement de Facebook : DEWEY, C., *98 personal data points that facebook uses to target ads to you*, The Washington Post, 19 août 2016, publié sur https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/?utm_term=.71534ba9eff5.

[199] Voir par analogie la jurisprudence concernant l'interception à grande échelle de télécommunications par les services de sécurité allemands, citée à la Cour européenne des droits de l'homme, le 26 juin 2015, Weber & Saravia c. Allemagne.

[200] Voir la recommandation n° 05/2012 de la Commission du 11 avril 2012, publiée à l'adresse suivante https://www.privacycommission.be/sites/privacycommission/files/documents/recommandation_05_2012_1.pdf.

[201] Voir le recours aux gestionnaires de réseaux de distribution dans la lutte contre la fraude sociale et la fraude à l'énergie.

[202] Voir les considérants 38 et 75 du RGPD,
pour autant que ce soit prévu dans le droit national.

[203] Article 36.5 du RGPD.

[203] En application de l'article 36.5 du RGPD, pour autant que ce soit prévu dans le droit national.

scientifique), plutôt qu'une analyse de données par des tiers qui peuvent être établis dans des pays sans niveau adéquat de protection des données[204] et/ou peuvent avoir des intérêts financiers dans la vente de solutions d'analyse de données ;

- renforcer le rôle des contrôleurs indépendants. En cas de fortes concentrations de données multinationales (fusions dans le secteur numérique,) le CEPD a avancé l'idée d'une collaboration renforcée entre les contrôleurs en protection des données, en protection des consommateurs et en compétition dans ce qu'on appelle la "Digital Clearing House"[205] ;
- pour la "predictive policing", on pourrait par exemple prévoir explicitement que l'Organe de Contrôle de la Gestion de l'Information Policière (COC) se voie attribuer une mission spécifique de contrôle afin d'évaluer l'opportunité et l'efficacité du projet i-Police de la Police fédérale [206] . L'Institut belge des services postaux et des télécommunications (IBPT) pourrait en outre veiller à ce que le monitoring par les opérateurs de télécommunications du trafic Internet des utilisateurs ne soit pas utilisé pour de la surveillance de prospection non autorisée de clients alors qu'il conviendrait d'adopter un principe de gestion neutre du réseau ;
- un renforcement de la position juridique d'ONG et d'organisations de droits civiques, via la confirmation légale de la possibilité d'une représentation collective des personnes concernées[207] afin d'introduire une plainte au nom de la personne concernée ou d'exercer en son nom les droits visés aux articles 77, 78 et 79 du RGPD ;
- une capacité et une expertise techniques et statistiques renforcées au sein des contrôleurs dont la Commission vie privée, le Comité P, le Comité R, ... ;
- l'intervention d'un Comité éthique.

Le RGPD instaure, comme on le sait, la fonction de délégué à la protection des données (articles 37 et suivants). Dans le cadre d'un contrôle et d'une surveillance renforcés des analyses de big data, ce délégué devra veiller à l'efficacité de ces garanties.

Autre constat : le contrôle des analyses de big data est trop souvent réparti de manière déséquilibrée et donc pas sur toutes les phases du big data qui ont été énumérées ci-avant. Pour le secteur public, la Commission a déjà attiré l'attention précédemment[208] sur la nécessité de prévoir des séparations entre la collecte des données et la transmission effective des résultats de l'analyse de ces données (phase de l'utilisation des données par l'inspection sociale, les gestionnaires de réseaux de distribution, ...).

---

[204] Voir les articles parus en septembre 2016 concernant la possibilité de participation d'une entreprise publique chinoise dans EANDIS et la page 15 de l'avis du VREG (Vlaamse Regulator van de Elektriciteits- en Gasmarkt, régulateur flamand du marché de l'électricité et du gaz) du 8 avril 2015 concernant un projet d'arrêté du Gouvernement flamand modifiant l'arrêté relatif à l'énergie du 19 novembre 2010 en ce qui concerne l'installation de compteurs intelligents, publié sur http://www.vreg.be/sites/default/files/document/adv-2015-03_ontwerp_van_besluit_uitrol_slimme_meters.pdf.

[205] CEPD, Avis n° 08/2016 on coherent enforcement of fundamental rights in the age of big data, 23 septembre 2016, publié sur https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Events/16-09-23_BigData_opinion_EN.pdf.

[206] MEEUS, R., Longread: *Hoe iPolice de natie veiliger maakt*, 24 juin 2015, http://datanews.knack.be/ict/nieuws/longread-hoe-ipolice-de-natie-veiliger-maakt/article-longread-720899.html ; Ponciau, L, *Les détails du projet iPolice dévoilés*, Le Soir, 17 septembre 2016.

[207] L'article 5.2 du RGPD prévoit une possibilité pour la Belgique de prévoir une telle mesure.

[208] Point 39 de l'avis n° 25/2016 de la Commission du 8 juin 2016 relatif à la fraude à l'énergie.

✓ **Recommandation 31**

Le contrôle doit porter sur l'ensemble de la chaîne de valeur du big data (voir ci-avant : collecte, stockage et préparation, analyse et utilisation) et pas uniquement sur la collecte initiale et l'utilisation ultérieure des données.

S. Sécurité des données à caractère personnel et violations de données à caractère personnel - approche basée sur le risque en vertu du RGPD

Plus grand est le nombre de données à caractère personnel et de traitements utilisés dans un projet de big data, plus grands sont les risques et les conséquences si une violation liée à la sécurité de ces données à caractère personnel[209] se présente (par exemple un abus de données par le personnel d'un responsable ou du sous-traitant, le vol de données qui sont ensuite diffusées sur le dark web).

Les responsables d'analyses de big data doivent dès lors à tout moment être attentifs à la prévention via des mesures de sécurité techniques, procédurales et organisationnelles adéquates et efficaces[210].

Ces mesures devraient tenir compte de la nature, de la portée, du contexte et des finalités du traitement ainsi que du risque que celui-ci présente pour les droits et libertés des personnes physiques. Le principe de la responsabilité ("accountability") du responsable du traitement va donc également de pair avec une approche basée sur les risques ("risk based approach")[211].

Une bonne méthode de gestion des risques constituera la base d'une sécurité adéquate et efficace[212], la notification de violations de données à caractère personnel à la Commission et à la personne concernée[213] et la réalisation d'une analyse d'impact relative à la protection des données[214].

✓ **Recommandation 32**

La Commission attire l'attention sur sa précédente recommandation du 21 janvier 2013 concernant la prévention des violations de données à caractère

---

[209] Cette notion est définie par le RGDP en son article 4, point 12 comme suit : "*une violation de la sécurité entraînant, de manière accidentelle ou illicite, la destruction, la perte, l'altération, la divulgation non autorisée de données à caractère personnel transmises, conservées ou traitées d'une autre manière, ou l'accès non autorisé à de telles données*"

[210] Considérant 74 du RGPD, par exemple la séparation des rôles fonctionnels dans le cadre de la gestion des données, du cryptage et de la gestion des accès.

[211] Voir WP 218, Déclaration du Groupe 29 du 30 mai 2014 sur le rôle d'une approche basée sur les risques dans des cadres juridiques de protection des données, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

[212] Article 32 du RGPD.

[213] Articles 33 et 34 du RGPD.

[214] Voir également les considérants 76 et 77 et les articles 32, 33 et 34 du RGPD. L'analyse du risque est au cœur du RGPD et particulièrement en lien avec la notification des brèches de sécurité. Le risque à prendre en compte n'est pas le risque lié à l'entreprise mais plutôt pour les droits et libertés des personnes concernées.

personnel[215].

✓ Recommandation 33

Les responsables et les sous-traitants impliqués dans des projets de big data doivent continuellement évaluer et gérer[216] les risques pour les droits et libertés des personnes concernées et les documenter en interne[217].

[215] CPVP, Recommandation d'initiative n° 01/2013 du 21 janvier 2013 *relative aux mesures de sécurité à respecter afin de prévenir les fuites de données (CO-A-2013-001)*, publiée sur https://www.privacycommission.be/sites/privacycommission/files/documents/recommandation_01_2013.pdf.

[216] Délimitation des critères de risques, pondération des critères et évaluation des risques (évaluation du risque et importance de l'impact).

[217] Article 30 du RGPD (registre des activités de traitement), article 33.5. du RGPD (pour toutes les violations de données à caractère personnel). Voir WP 218, Déclaration du Groupe 29 du 30 mai 2014 sur le rôle d'une approche basée sur les risques dans des cadres juridiques de protection des données, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

Lexique

Agrégation :	Le remplacement de groupes d'observations ou d'objets par des informations récapitulatives.
Algorithme :	Une série ordonnée d'étapes univoques exécutables décrivant un processus fini. Il s'agit en d'autres termes d'une série d'instructions permettant aux ordinateurs de solutionner un problème ou d'obtenir un certain résultat.
Distorsion de données :	Le phénomène selon lequel les données ne sont pas systématiquement représentatives de la population que l'on étudie. Il est dû à des erreurs systématiques lors de la création de l'ensemble de données. Il en résulte que les estimations que l'on obtient en utilisant ces données s'écarteront systématiquement de la valeur réelle que l'on souhaite calculer.
Data Mining :	Le data mining est le processus d'analyse de grandes quantités de données brutes, dans une recherche d'informations et de connaissances exploitables sous la forme de modèles et de relations.
Modèle (mathématique) :	Un modèle mathématique décrit (ou pressent) la relation entre différentes variables (par exemple entre des variables d'input et des variables d'output où une variable d'output quantifie par exemple le risque de fraude) sous une forme mathématique.
Surapprentissage :	La modélisation d'une variation aléatoire dans un ensemble d'apprentissage qui n'a rien à voir avec la relation sous-jacente (entre les variables) que l'on entend capter, impliquant un modèle mathématique qui fait de moins bonnes prévisions pour de nouvelles données inconnues (ensemble de test).
Predictive policing :	Le terme renvoie à l'utilisation de techniques mathématiques, de prédiction et d'analyse dans le maintien de l'ordre afin de détecter une potentielle activité criminelle. Dans ce cadre, il peut s'agir de méthodes visant à prévoir les délits, de méthodes visant à prévoir la criminalité ou de méthodes visant à prévoir l'identité d'auteurs ou la victimisation de la criminalité.
Pseudonymisation :	Voir la définition à l'article 4, 5) du RGPD : "pseudonymisation", le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.
Variable quasi ID :	Les caractéristiques qui sont utilisées pour délimiter les groupes dans des données agrégées mais qui, en combinaison, pourraient aussi être utilisées en vue d'identifier un individu. Des exemples sont le pays du domicile, le code postal, le sexe, l'âge ou la date de naissance. Le risque d'identification dépend du nombre et de la nature de ce type de variables dans les données et des connaissances a priori de celui qui tente de procéder à l'identification.
Individualisation :	Le suivi individuel de personnes physiques sans connaître nécessairement leur identité civile ou numérique.
Analyse de risques Small Cells :	Une analyse des risques qui examine le risque d'une (ré)identification dans des données agrégées.
Unicité :	Le risque de réidentification d'un individu dans un ensemble de données, compte tenu d'une quantité déterminée d'informations dont on dispose déjà concernant l'individu, provenant d'autres sources

Sources

"Wetenschappelijke Raad voor het Regeringsbeleid" (Conseil scientifique de la politique gouvernementale aux Pays-Bas), *Big Data in een vrije en veilige samenleving*, University Press Amsterdam, publié à l'adresse http://www.wrr.nl/fileadmin/nl/publicaties/PDF-Rapporten/rapport_95_Big_Data_in_een_vrije_en_veilige_samenleving.pdf.

Wetenschappelijke Raad voor het Regeringsbeleid, Synopsis van WRR-rapport, http://www.wrr.nl/fileadmin/nl/publicaties/PDF-samenvattingen/Synopsis_R95_Big_Data_in_vrije_en_veilige_samenleving.pdf.

Lignes directrices sur la protection des personnes à l'égard du traitement des données à caractère personnel à l'ère des mégadonnées, Comité consultatif de la convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, T-PD(2017)01, 23 janvier 2017, publié à l'adresse

- https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebf2