Le paradoxe de la protection des données personnelles à l'heure de la libre circulation des informations

Nadia Hassani

Terminal

Technologie de l'information, culture & société

124 | 2019
Vivre dans un monde sous algorithmes
Vivre dans un monde sous algorithmes

Le paradoxe de la protection des données personnelles à l'heure de la libre circulation des informations

Quel cadre éthique offre le RGPD aux data scientists ?

The paradox of personal data protection in the age of the free flow of information : which ethical framework the GDPR offers to the data scientists ?

Nadia Hassani
https://doi.org/10.4000/terminal.4040

Résumés

Français English
Le 25 mai 2018, le Règlement général pour la protection des données personnelles (RGPD) est entré en vigueur dans toute l'Union européenne. Celui-ci impose aux entreprises de mettre en conformité les conditions générales d'utilisation de leurs services numériques et leurs méthodes de collecte et de traitement des données. Notre cherchons ici à étudier par quels dispositifs info-communicationnels les responsables de l'exploitation des données massives que sont les data scientists ont pu être sensibilisés et formés pour évoluer dans ce nouveau cadre réglementaire européen. Dans une démarche empirique et exploratoire s'inscrivant dans le champ des Sciences de l'Information et de la Communication, et à la lumière d'une revue de la littérature scientifique et médiatique, notre proposition de communication présente les résultats d'une enquête quantitative et qualitative menée en ligne auprès d'un panel représentatif de data scientists français sélectionnés sur le réseau social numérique professionnel LinkedIn. Il met en lumière les difficultés rencontrées par ces professionnels chargés désormais d'assurer un traitement éthique des données massives, tout en participant à leur libre circulation et à leur marchandisation.

On May 25th, 2018, the General Data Protection Regulation (GDPR) became effective

throughout the European Union. It requires companies to bring the terms and conditions of their digital services into compliance with their data collection and processing methods. The purpose of our research is to study by which information-communication devices the data scientists have been made aware and trained to evolve in this new European regulatory framework. In an empirical and exploratory approach in the field of Information and Communication Sciences, and in the light of a review of the scientific and media literature, this article presents the results of a quantitative and qualitative survey conducted online of a representative panel of French data scientists selected on the professional social network LinkedIn. It highlights the challenges encountered by these professionals now responsible for ensuring an ethical treatment of massive data, while participating in their free movement and their commodification.

Entrées d'index

Mots-clés: RGPD, data scientists, big data, données personnelles, éthique
Keywords: GDPR, data scientists, big data, personal data, ethics

Texte intégral

1 En moins de deux décennies, les TIC (Technologies de l'Information et de la Communication) sont devenues omniprésentes dans nos vies personnelles et professionnelles. En 2017, la part des ménages français ayant accès à Internet était de 86 % (Eurostat, 2018) alors qu'elle n'était que de 14,3 % en 2000. Aujourd'hui, la plupart de nos tâches quotidiennes, des plus élémentaires aux plus complexes, sont médiées par des dispositifs info-communicationnels (SMS, courriels, applications, réseaux sociaux ou objets connectés) dont l'essor fait figure de progrès technique indispensable, moteur de développement économique, social et culturel. À tel point que le terme « illectronisme » a fait son entrée en 2018 dans le dictionnaire, désignant la difficulté à utiliser l'Internet dans la vie de tous les jours. Alors que la fracture numérique – concept largement étudié dans le champ des Sciences humaines depuis la fin des années 1990 et désignant l'inaccessibilité des dispositifs de TIC (et plus spécifiquement à la téléphonie mobile, à l'outil informatique et à Internet) – semble se réduire à mesure du déploiement des réseaux haut débit et de l'évolution du taux d'équipement des français en smartphones (75 % en 2018 selon le Crédoc), un nouvel enjeu émerge dans ce contexte d'hyperconnectivité : la protection des données personnelles à l'heure des mégadonnées (ou big data). Plus de 10 ans après l'apparition de ce terme dans la littérature scientifique[1], l'exploitation des « données massives » représente actuellement l'un des piliers d'une économie numérique mondialisée et largement dominée par les « big tech » (terme désignant les GAFAM - Google, Apple, Facebook, Amazon et Microsoft – ainsi que leurs homologues internationaux tels que Baïdu, Alibaba, Samsung ou encore Vkontakte). Les prédictions annoncent un marché international de plus de 200 milliards de dollars d'ici 2020, ainsi qu'une véritable explosion des données stockées qui atteindront 175 Zo (soit $175x10^{21}$ octets) dès 2025 – soit 5,3 fois plus qu'aujourd'hui[2] – nous conduisant à très haut débit de l'ère du big data à celle du huge data.

Scandales, cybercriminalité, fuites de données massives : le revers de la donnée

2 Alors que la donnée est considérée comme « le pétrole du 21ème siècle »

(Babinet, 2014), des scandales de grande ampleur entachent régulièrement l'image de ces géants de l'industrie numérique. Chaque fois, l'utilisation abusive des données personnelles des utilisateurs confirme le danger que représente la non-régulation de leur exploitation, la vigilance des médias et la préoccupation du grand public à cet égard. Le 17 mars 2018, la société britannique Cambridge Analytica se retrouve au cœur d'une affaire révélée par le New York Times : celle-ci concerne l'exploitation de données personnelles d'environ 87 millions d'utilisateurs du réseau social numérique Facebook à des fins politiques. Le 30 août 2018, Bloomberg dévoilait un accord secret entre Google et MasterCard aux États-Unis : il concerne les clients MasterCard américains ayant un compte Gmail et ayant accepté la politique de suivi publicitaire dans les paramètres de leur compte Google. Ceux-ci ont fait l'objet d'un recoupement entre leurs données bancaires (achats effectués hors ligne, en magasins physiques) et leurs profils utilisateurs. Autre exemple datant celui-ci du 20 septembre 2018 : le Wall Street Journal de mettre au jour le fait que Google continue à autoriser des applications tierces à analyser et à partager les données personnelles des comptes Gmail, alors que le géant américain avait annoncé avoir mis un terme à cette pratique de ciblage publicitaire en 2017. Plus récemment encore, le Wall Street Journal révélait le 22 février 2019 un autre scandale concernant Facebook, accusé de récupérer des données personnelles sensibles (comme le poids, les cycles menstruels ou les phases d'ovulation) issues d'applications tierces populaires, et ce même si l'usager ne possède pas de compte Facebook[3]. Nous voyons, à travers ces scandales à répétition, que des fragments de notre identité numérique, voire de notre intimité numérique (Simonin, 2015) circulent, sont vendus et exploités à des fins de ciblage publicitaire, le plus souvent à notre insu. Nos données personnelles font aujourd'hui et plus que jamais l'objet de convoitises à caractère marchand, à l'origine de la cybercriminalité et de fuites de données massives. Le 5 décembre 2018, le Quai d'Orsay a été victime d'un piratage de sa base de données « Ariane », plateforme qui permet de préparer un déplacement à l'étranger, de s'enregistrer en ligne et d'obtenir des informations relatives à la sécurité du pays de destination. Les données personnelles de 540 563 Français ont été dérobées par des cybercriminels, comprenant notamment des données intimes telles que la personne à contacter en cas d'urgence. Le 14 février 2019, la presse spécialisée révélait que les données de 16 applications regroupant au total 620 millions de comptes étaient en vente sur le « darknet », cet Internet clandestin constitué d'un ensemble de réseaux de type « friend to friend » conçus pour assurer l'anonymat de leurs utilisateurs (Jomni, 2018). Plus récemment encore, c'était au tour de l'Icann[4], le mainteneur de l'annuaire central de l'Internet, d'être la cible d'une cyberattaque le 25 février 2019, faisant craindre une fuite de données d'une ampleur inédite.

Vers une gestion plus éthique de la donnée en Europe ?

3 Pour répondre aux préoccupations grandissantes et légitimes de ses citoyens, l'Union européenne s'est dotée le 25 mai 2018 d'un cadre législatif avec l'entrée en vigueur du RGPD : le Règlement général pour la protection des données personnelles. Adopté deux ans plus tôt par le Parlement européen, il impose aux entreprises de mettre en conformité les conditions générales d'utilisation de leurs services numériques et leurs méthodes de collecte et de traitement des données. Alors que ce nouveau cadre réglementaire est censé renforcer la sécurité de nos données personnelles, le RGPD a également vocation à autoriser leur libre

circulation. Quatre mois après son entrée en application, la Cnil a publié un rapport datant du 25 septembre 2018 dans lequel elle déclare avoir reçu 600 notifications de violations de données concernant environ 15 millions de personnes. La commission a également recueilli 3767 plaintes – soit une augmentation de 64 % par rapport à la même période en 2017 – signe que les citoyens européens se sont bien appropriés le nouveau cadre proposé par le RGPD[5].

4 La dernière décennie a vu apparaître l'avènement de l'ère des mégadonnées et de nouveaux métiers comme les architectes de la donnée, les ingénieurs big data, les analystes de la donnée, les délégués à la protection des données (DPD, ou DPO en anglais pour « Data Protection Officers ») ou encore les « data scientists », ces experts de la science de la donnée auxquels nous avons souhaité nous intéresser plus particulièrement dans le cadre de ce travail de recherche. En une décennie, la fonction s'est considérablement développée au sein des entreprises, en France comme à l'étranger. Et c'est dans le cadre d'un autre travail de recherche portant sur le bonheur au travail que nous avons découvert que le métier de data scientist arrivait chaque année en tête du classement américain GlassDoor, et ce depuis 2016. Ce site Web qui permet aux employés et anciens employés d'entreprises du monde entier d'évaluer anonymement leur environnement de travail le classe depuis quatre années consécutives comme « the best job » (le meilleur emploi), que ce soit en matière de rémunération, d'opportunités de carrière ou de satisfaction au travail, avec un score de 4,3/5 pour ce dernier facteur[6]. Ceci nous a particulièrement motivé pour en savoir plus sur cette profession et pour comprendre comment, dans un contexte où la cybersécurité est devenue un enjeu international, ceux qui exercent ce job de rêve en France ont accueilli la mise en place du RGPD, comment ils y ont été sensibilisés ou formés, et si pour eux cette nouvelle réglementation européenne participe à un traitement plus «thique des données personnelles récoltées.

Terrain de recherche, constitution de l'échantillon et méthodologie

5 Pour répondre à cette problématique, nous avons initié notre travail de recherche en septembre 2018 en constituant dans un premier temps un corpus composé d'articles de presse, de sites Web de référence, d'interviews, de dossiers thématiques, ou de vidéos qui nous ont permis de nous familiariser avec ce métier. Une étude approfondie de ce corpus nous a permis de découvrir à quel point les data scientists occupent une position stratégique dans les entreprises à l'heure de l'économie numérique. Nous pourrions les définir comme les responsables de la gestion et de l'analyse de « données massives » récoltées de sources numériques très hétérogènes (Web, téléphone mobile, réseaux sociaux, transactions bancaires, etc.) qu'ils doivent exploiter afin de leur donner du sens en les croisant avec les données propres à leur entreprise (données fournisseurs, clients, marché, etc.). Leurs compétences de haut niveau, leur polyvalence et leur vision transverse leur permettent d'extraire des informations d'ordre stratégique ou opérationnel facilitant la prise de décisions des entreprises, quel que soit leur secteur d'activité.

6 Afin d'étudier plus concrètement notre objet de recherche et tenter de répondre à notre problématique, nous avons souhaité adopter une démarche empirique dans le but de produire des connaissances inédites dans le champ des sciences de l'information et de la communication, et faire ainsi émerger des éléments de réflexion s'appuyant sur l'analyse des résultats d'une analyse qualitative et d'une enquête quantitative que nous avons menées successivement.

Méthodologie de notre analyse quantitative

7 Tout d'abord, nous avons réalisé une analyse quantitative menée en ligne sur le réseau social professionnel LinkedIn que nous pouvons définir comme un artefact communicationnel où les interactions sociales sont médiatisées. Créé en 2003, LinkedIn est un réseau social à visée professionnelle qualifié également de « site de réseautage » (Stenger et Coutant, 2013). De par ses dernières évolutions, le dispositif sociotechnique qu'est LinkedIn peut désormais être considéré comme une véritable plateforme de production de contenus numériques textuels (articles, messages privés, commentaires) et d'interactions sociales (invitations, recommandations, « j'aime », « félicitations », etc.) dont la diffusion peut être aussi bien publique que restreinte comme dans le cas des groupes privés par exemple. Au début de notre enquête, le moteur de recherche LinkedIn référençait plus de 680 000 profils de data scientists dans le monde ; en filtrant les résultats de notre requête sur la France, la plateforme en affichait précisément 21 373. Pour réaliser notre enquête quantitative sur la base d'un échantillon représentatif, et bien que la méthode que nous avons retenue nécessite un temps et un investissement importants pour constituer un corpus significatif, nous avons choisi de nous concentrer sur les 500 premiers profils de data scientists français proposés par le moteur de recherche LinkedIn. Après avoir pris soin d'exclure de notre échantillon les personnes en recherche d'emploi, en stage ou en formation, ainsi que les profils mal ou non renseignés, et ceux qui nous ont semblé suspects (c'est à dire n'ayant pas de formation ou d'expérience correspondant à l'intitulé du poste présenté sur leur profil), nous avons ensuite analysé les 500 profils retenus, un à un, en fonction des 7 critères suivants : le genre (afin de déterminer la proportion hommes-femmes au sein de cette profession), leur situation géographique, le niveau d'études, le type de cursus (école d'ingénieur, université, etc.), le domaine de formation, le nombre d'années d'expérience et enfin le secteur d'activité actuel. Nous avons classé, trié et compilé les informations récoltées en fonction de ces 7 critères dans un tableur Excel qui a généré des statistiques précises sur la base de cet échantillon de 500 profils.

Méthodologie de notre enquête qualitative

8 Les résultats de notre analyse quantitative que nous présenterons ci-après nous ont permis de mieux cerner le profil des data scientists français, et de préparer ainsi une grille d'entretiens semi-directifs concentrée essentiellement sur la gestion éthique des données personnelles dans le nouveau contexte règlementaire européen du RGPD, en évitant les questions d'ordre général.

9 Pour constituer notre nouvel échantillon et identifier les data scientists à qui nous avons souhaité proposer un entretien, nous nous sommes concentrés sur les profils identifiés lors de notre analyse quantitative qui nous ont semblé être les plus expérimentés et les plus actifs sur le réseau social professionnel. Une expérience d'au moins trois ans à un poste de data scientist, un réseau de plus de 500 relations, un profil renseigné avec le détail des compétences et des outils maîtrisés, ainsi que la publication régulière d'articles sur leur fil d'actualité nous ont semblé être des critères intéressants pour présélectionner 48 profils LinkedIn afin de constituer un nouvel échantillon en adéquation avec les statistiques obtenues lors de notre analyse quantitative. Nous les avons ensuite contactés via l'outil de messagerie LinkedIn, ce qui nous a permis de leur présenter succinctement notre projet de recherche et de leur demander s'ils accepteraient de répondre à nos questions dans le cadre d'un court entretien téléphonique, en précisant que celui-ci serait enregistré à des fins d'analyse et que leur anonymat

serait garanti. Sur 48 demandes, 31 data scientists ont répondu à notre message et 17 ont finalement été interviewés : soit par téléphone (11 personnes), soit via l'application Skype (6 personnes), ce qui représente un taux d'acceptation de seulement 34,5 %. Ce chiffre est cependant à relativiser du fait que nous ayons cherché à maintenir une proportion hommes-femmes proche de celle constatée à l'issue de notre analyse quantitative des 500 profils : afin que nos entretiens soient réalisés auprès d'un échantillon aussi représentatif que possible, nous n'avons pas retenu 4 personnes nous ayant préalablement donné leur accord de principe (le taux d'acceptation aurait donc été en réalité de 45,8 %).

10 Ces 17 entretiens semi-directifs ont été menés entre le 9 janvier et le 12 février 2019 ; ils ont duré en moyenne 26 minutes, ce qui peut sembler relativement court mais qui s'explique par une grille d'entretien resserrée autour de 3 questions principales :

- Comment avez-vous été sensibilisé(e) et formé(e) au RGPD ?
- Comment la mise en place du RGPD a-t-elle modifié votre processus de collaboration et les relations que vous entretenez avec vos différents interlocuteurs ?
- Selon vous, en quoi le RGPD est-il propice à l'émergence d'une gestion éthique des données massives ?

11 L'analyse quantitative nous ayant permis d'obtenir de nombreuses informations sociodémographiques, nous avons ainsi pu réduire la durée des entretiens et aborder plus directement les questions en lien direct avec notre problématique de recherche. Nous avons ainsi pu collecter près de neuf heures d'enregistrement audio grâce à l'application « dictaphone » de notre téléphone portable. Nous avons ensuite retranscrit ces données primaires grâce à l'application en ligne « Google Docs » qui propose un service de saisie vocale couplé à un enregistrement automatique et en temps réel du texte généré. Une fois cette fonctionnalité activée, nous avons procédé à la lecture de chacun des 17 enregistrements après avoir mis notre téléphone en haut parleur, afin que chaque piste puisse être enregistrée distinctement via le microphone de notre ordinateur. Nous avons ensuite regroupé les réponses pour chacune des trois questions posées. Celles-ci étant délibérément ouvertes, nous avons fait le choix de lister les données qualitatives (noms, notions ou thématiques) contenues dans chaque réponse afin de les traduire dans un second temps en données quantitatives pour pouvoir illustrer les propos qui nous ont été rapportés à l'aide de données chiffrées. Une analyse lexicométrique des données récoltées nous a permis de mieux appréhender leur subjectivité. En effet, la quantité et la diversité des données récoltées nous ont conduit à élaborer dans un premier temps un dictionnaire des thèmes empirique nous permettant d'analyser les contenus thématiques émergeant de chaque verbatim. Pour élaborer cette codification, nous avons quantifié manuellement les fragments signifiants (mots, concepts, expressions) présents dans chacune des réponses produites par les data scientists pour les classer ensuite dans différents thèmes empiriques et obtenir une distribution significative des éléments de discours obtenus. Nous les avons ensuite classés et triés dans une base de données Excel afin de générer des statistiques correspondant à la fréquence de ces données qualitatives préalablement codifiées et catégorisées. L'intérêt d'une telle analyse de contenu réside dans le fait qu'elle « permet, lorsqu'elle porte sur un matériau riche et pénétrant, de satisfaire harmonieusement aux exigences de la rigueur méthodologique et de la profondeur inventive qui ne sont pas toujours facilement conciliables » (Quivy et Van Campenhoudt, 2011). En appliquant cette méthodologie, et après une phase de relecture, de correction des erreurs de retranscription et de suppression des

passages inutiles dans les textes générés par ce procédé, nous avons obtenu un corpus total de 26 128 mots sur lequel nous avons effectué un travail d'analyse dont les résultats sont présentés ci-après.

Présentation des résultats

Qui sont les data scientists français ?

12 Le premier enseignement que nous révèle l'analyse de cet échantillon de 500 profils LinkedIn est la forte masculinisation de ce poste en 2019 : 77,6 % des data scientists sont des hommes (388 profils), alors que les femmes sont représentées à hauteur de 22,4 % (112 profils). Ils sont très majoritairement basés en région parisienne (63,1 %), dans la région de Montpellier (10,8 %) ou en région lyonnaise (8,6 %). Quel que soit le genre ou leur situation géographique, nous avons constaté que les data scientists ont des profils hautement qualifiés : 83 % de ces professionnels possèdent un diplôme d'ingénieur ou de niveau équivalent, et près de 8 % d'entre eux sont titulaires d'un doctorat. La tableau suivant détaille les différents types de formations suivies par les data scientists, sachant qu'une part importante d'entre eux ont effectué un double cursus :

Tableau 1 : représentativité des cursus suivis par les data scientists

Cursus suivis	Pourcentages des profils analysés
Université	48 %
Ecole d'ingénieur	38,70 %
Polytechnique	7,90 %
Ecole de commerce ou management	2,66 %
Ecole Normale Supérieure	2,63 %

13 Il est intéressant de constater la part très importante de l'enseignement universitaire dans les cursus des data scientists : l'université participe en effet à la formation de ces cadres de haut niveau pour près de la moitié d'entre eux. Nous nous sommes également intéressés aux domaines de formation qu'ont suivi ces scientifiques de la donnée, et nos résultats d'analyse sont présentés dans le tableau suivant :

Tableau 2 : les domaines de formations suivis par les data scientists

Domaines de formation	Pourcentages des profils analysés
Mathématiques et statistiques	33,50 %
Sciences informatiques	20,90 %
Big Data, Sciences de la donnée	15,00 %
Economie / Gestion / Finance	10,20 %
Sciences physiques	3,90 %
Electronique / Telecommunications	3,90 %
Management	3 %

Intelligence artificielle	2,70 %
Marketing / communication	2,10 %
Biologie	1,50 %
Sciences appliquées	1,20 %
Chimie	0,90 %
Neurosciences	0,60 %
Agronomie / Environnement	0,60 %

14 Nous constatons la dimension scientifique de cette profession puisque plus de la moitié des data scientists ont suivi un parcours dans l'enseignement supérieur dans les domaines des mathématiques et des statistiques (33,5 %) ou de l'informatique (20,9 %). Notons également que 15 % d'entre eux sont titulaires d'un diplôme spécialisé dans les sciences de la donnée, et que nos résultats révèlent également que plus d'un data scientist sur dix (10,2 %) est issu d'une formation supérieure dans le domaine de l'économie ou de la finance. Intéressons-nous maintenant aux secteurs d'activité dans lesquels évoluent les data scientists exerçant en France :

Tableau 3 : les secteurs qui recrutent les data scientists

Secteurs d'activité	Pourcentages des profils analysés
Cabinets et agences conseil spécialisés	31,10 %
Communication / Marketing / Publicité	13,90 %
Banque / Finance / Assurances	11,60 %
IT / Informatique	9,80 %
Sites Web / E-commerce	5,90 %
Transports /Mobilité	4,40 %
Internet / Télécommunications	3,80 %
Luxe / Cosmétiques	3,10 %
Organismes publics	2,60 %
Grande distribution	2,30 %
Énergie	2,20 %
Industrie médico-pharmaceutique	2,10 %
Médias	1,70 %
Industrie	1,20 %
Laboratoires / centres de recherches	1,10 %
Indépendant / Freelance	1 %
Environnement	0,90 %
Tourisme / Loisirs	0,60 %

Enseignement supérieur	0,50 %
Ressources humaines	0,20 %

15 Nous constatons que si la plupart des secteurs d'activité semble avoir amorcé une transformation numérique, la science de la donnée est une expertise que beaucoup d'entreprises externalisent en faisant appel à des agences spécialisées en science de la donnée (31,10 %), des agences de communication ou de marketing (13,9 %) ou des entreprises du domaine de l'informatique (9,8 %). Ces trois secteurs représentent à eux seuls plus de la moitié (54,8 %) des entreprises dans lesquelles évoluent les data scientists français. Notons toutefois que le secteur de la finance arrive en troisième position, accueillant plus de un data scientist sur dix (11,6 %). Enfin, le dernier critère que nous avons souhaité étudier dans le cadre de cette analyse quantitative est le nombre d'années d'expérience affiché sur leurs profils LinkedIn, lequel est détaillé dans le tableau ci-dessous :

Tableau 4 : le niveau d'expérience des data scientists français

Nombre d'années d'expérience	Pourcentages des profils analysés
< 1 an	13 %
1 à 2 ans	27,80 %
2 à 3 ans	18,34 %
3 à 4 ans	17,16 %
4 à 5 ans	12,42 %
5 à 6 ans	5,91 %
6 à 7 ans	1,77 %
7 à 8 ans	1,77 %
8 à 9 ans	1,18 %
9 à 10 ans	0,60 %

16 Nous voyons ici clairement que si la profession n'est pas nouvelle (certains profils revendiquent en effet une expérience de dix ans à ce poste), près de 60 % (59,14 %) des data scientists ont moins de trois ans d'expérience, ce qui signifie que l'industrie de la donnée a suscité ces trois dernières années de réelles vocations, conduisant les établissements de l'enseignement supérieur à s'adapter pour proposer des formations spécialisées dans l'analyse et la gestion de la donnée afin de répondre aux besoins actuels des entreprises, tous secteurs confondus.

17 Au regard de l'analyse quantitative des 500 profils que nous avons effectuée, nous pourrions donc dresser un premier profil type du data scientist français : un homme basé en région parisienne ayant fait de hautes études scientifiques (en mathématiques ou informatique), et travaillant dans une agence spécialisée depuis moins de trois ans. Nous avons cependant souhaité aller au-delà des chiffres et des données récoltées grâce à une enquête qualitative pour permettre aux data scientists qui ont accepté de nous répondre d'apporter leur éclairage sur l'impact qu'a eu l'entrée en vigueur du RGPD dans leur métier.

Que pensent les data scientists de l'apport

éthique du RGPD ?

18 Nous avons cherché à étudier par quels dispositifs info-communicationnels les responsables de l'exploitation des données massives que sont les data scientists ont pu être sensibilisés et formés pour évoluer dans ce nouveau cadre réglementaire européen.

Résultats d'analyse des réponses à la question « Comment avez-vous été sensibilisé(e) et formé(e) au RGPD ? »

19 L'analyse des entretiens semi-directifs a permis de mettre en avant deux grandes catégories de réponses : les démarches initiées par l'employeur et les initiatives personnelles.

Tableau 5 : démarches « employeur »

Formations avec cabinets avocats et juristes spécialisés	39,40 %
Réunions internes (Manager, Lead Data Scientist)	28,70 %
Séminaires et conférences	21,60 %
Intranet / Réseaux sociaux professionnels	9,10 %
Mails et messageries internes	1,20 %

Tableau 6 : initiatives personnelles

Presse et sites web spécialisés	28,10 %
Guides PIA	21,60 %
Webinars et vidéos en ligne	18,80 %
Veille réseaux sociaux	15,90 %
Livres	10,70 %
Newsletters	4,90 %

20 Nous pouvons constater ici la grande hétérogénéité des dispositifs info-communicationnels mis en œuvre au sein des entreprises pour former les data scientists interrogés. Leur apprentissage du RGPD est résolument « hybride » et repose sur un équilibre entre formation interne (en grande majorité assurée par des cabinets juridiques experts en RGPD, cité dans 39,4 % des cas) et une veille documentaire personnelle effectuée essentiellement en ligne sur des sites spécialisés (28,10 %) ou à partir de « guides PIA » (21,60 %) qui sont des documents de référence présentant les bonnes pratiques en matière de traitement des données personnelles et de respect de la protection des données à caractère personnel ou « privacy », anglicisme fréquemment employé par les data scientists que nous avons interviewés.

Résultats d'analyse des réponses à la question « Comment la mise en place du RGPD a-t-elle modifié votre processus de collaboration et les relations

que vous entretenez avec vos différents interlocuteurs ? »

Tableau 7 : impact du RGPD sur les procédures et les relations

DPO	51,20 %
Polyvalence	21,10 %
Coopération	12,70 %
Complexification	8,30 %
Confiance	6,70 %

21 Le fait que la polyvalence soit une notion exprimée dans 21,1 % des cas est lié au fait que les répondants ont souligné que leur métier de data scientist est résolument transversal, de par ses dimensions techniques, communicationnelles voire commerciales qui les conduisent à être en lien direct et permanent avec une grande diversité d'interlocuteurs métiers. Une fois la donnée extraite, son rôle va être de lui donner du sens et de la mettre en forme (« data visualization ») afin de la rendre pertinente pour le client. Mais le terme qui est le plus présent dans les réponses que nous avons obtenues à cette deuxième question concerne un interlocuteur avec lequel les data scientists sont amenés à échanger régulièrement : le délégué à la protection des données (DPD), qu'il soit un prestataire externe ou membre de l'équipe projet à part entière. Depuis la mise en place du RGPD, le DPD est le référent qui permet aux data scientists de s'assurer que leurs projets sont bien conformes (« compliant ») à la réglementation en matière de protection des données personnelles. La présence de cet acteur est en lien direct avec les notions de coopération, de confiance, mais également de complexification des relations et des procédures comme l'explique très bien ce court extrait d'entretien :

22 « Quand je briefe le DPO, je dois faire en sorte qu'il comprenne précisément le périmètre et l'objectif du projet. Il doit avoir une vision claire des outils utilisés, du type de données à collecter et à traiter, du nombre de personnes et de partenaires sur le projet etc. C'est une étape aujourd'hui indispensable ; même si elle prend du temps, c'est lui qui va nous donner le feu vert pour avancer et nous garantir que le projet est "RGPD compliant" . »

23 La coopération apparaît comme une qualité émergente de la mise en place du RGPD pour plus d'un data scientist sur dix, que ce soit en interne (DPD) ou à l'externe avec les partenaires et sous-traitants qui doivent également être en conformité avec les nouvelles dispositions réglementaires.

Analyse des réponses à la question « Selon vous, en quoi le RGPD est-il propice à l'émergence d'une gestion éthique des données massives ? »

24 Cette troisième et dernière question de notre enquête qualitative nous amène au cœur de notre problématique de recherche. L'analyse des réponses obtenues nous a permis de mettre en avant deux grandes catégories d'arguments présentés dans les deux tableaux ci-dessous : la dimension éthique inhérente au RGPD telle que perçue par les data scientists que nous avons interrogés, et les limites que ceux-ci ont également exprimées dans le cadre de nos entretiens.

Tableau 8 : la dimension éthique du RGPD vue par les data scientists

Plus de sécurité	48,20 %

Plus grand respect de la vie privée	31,50 %
Prévention des dérives	17,40 %
Protection de dimension internationale	2,90 %

25 Nous voyons ici que le principal argument en faveur du RGPD aux yeux des data scientists est le fait que celui-ci participe à améliorer la sécurisation des données, celui-ci représentant près de la moitié des éléments signifiants avancés : « Pour être éthiques, les données doivent avant tout être sécurisées ! » a affirmé l'un des répondants à notre enquête. La prise en compte de la vie privée (31,6 %) dès le début et tout au long du projet de « data science » est également cité dans près d'un tiers des réponses collectées :

26 « Aujourd'hui on demande à nos clients de nous fournir des données anonymes, ce qui nous permet de travailler en "Privacy by design" (…) Le respect de la vie privée et la protection des données personnelles est une priorité, dès la conception du projet. »

27 Plus de 17 % des éléments de langage collectés nous montrent que le RGPD apparaît comme une protection permettant d'encadrer la collecte et le traitement de données dites sensibles en lien avec la santé (qu'elle soit mentale ou physique), les orientations sexuelles, les origines ethniques, les pratiques religieuses, voire même les émotions comme en atteste ce verbatim :

28 « Grâce à l'intelligence artificielle, il est déjà possible d'analyser l'état émotionnel des personnes. Je vous laisse imaginer l'intérêt que cela peut représenter dans le secteur du marketing ou de la publicité… Le RGPD impose que ces données sensibles collectées, par exemple via un système de vidéo-surveillance ou de reconnaissance faciale, aient un cycle de vie très court afin qu'elles ne puissent pas être archivées. »

29 Enfin, la dimension internationale a été soulignée par deux répondants qui nous ont précisé que « les partenaires étrangers doivent aussi se conformer au RGPD », et qu' « une entreprise française ou européenne devra toujours se conformer au RGPD, même si le client est basé à l'étranger ».

Tableau 9 : les limites du RGPD en matière d'éthique de la donnée

Responsabilisation	31,60 %
Quantité / Diversité	26,60 %
Multiplicité	19,20 %
Dissuasion	13,10 %
Irreversibilité	9,50 %

30 Ces derniers résultats d'analyse des discours recueillis révèlent la limite principale du RGPD telle que perçue par les data scientists : la quantité (26,6 %) et la diversité (19,2 %) des données produites et collectées chaque jour sont jugées incompatibles avec un traitement éthique du fait de l' « incapacité de l'utilisateur à pouvoir contrôler réellement toutes ses données ou à consentir à les mettre à disposition ». Face à ce constat, près d'un tiers des data scientists (31,6 %) considère que c'est aux professionnels de la donnée d'être les garants d'un traitement éthique des données personnelles, et qu'une réglementation européenne est « importante mais ne sera jamais suffisante » comme en atteste cet autre verbatim :

31 « Il n'y a pas de data éthique ou pas éthique en soi, ce sont les ingénieurs qui configurent les algorithmes, vont chercher la data et interprètent les données qui doivent faire preuve d'une certaine éthique. Le RGPD a le mérite d'exister, ce n'est

certainement pas suffisant, mais c'est un vrai référentiel pour toute la profession qui a déjà beaucoup à faire pour balayer devant sa porte ! »

32 Il est également intéressant de souligner que le RGPD leur paraît être un outil de dissuasion dans 13,1 % des cas : « Les pénalités de 4 % sur le chiffre d'affaire ? C'est avant tout pour faire peur aux fraudeurs ». Enfin, nous avons regroupé dans la catégorie irréversibilité tous les éléments de discours signifiant que « il n'y a pas de retour en arrière possible ». Les principaux arguments cités par les répondants sont que les algorithmes et les technologies du type « Machine Learning » ou « Natural Language Processing » progressent de manière « exponentielle », et que les « possibilités quasi infinies » qu'ils offrent génèrent « malheureusement » des « intérêts financiers plus importants que la morale ».

Limites de notre travail de recherche et conclusion

33 Utilisées conjointement, les analyses quantitatives et qualitatives nous ont permis d'objectiver les données récoltées lors de nos entretiens semi-directifs et de dresser un premier portrait des data scientists français. Mais l'une des premières limites que nous souhaitons exposer ici est le fait que notre enquête ne prenne pas en compte le cadre juridique préexistant. En effet, le RGPD fait suite à la directive européenne « 95/46/CE » et, dans le cas spécifique de la France, à la loi « Informatique et Libertés » datant de janvier 1978 (loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés). Par l'intermédiaire de notre questionnaire, il aurait été intéressant de demander aux data scientists ayant le plus d'expérience dans quelles mesures le RGPD est en lui-même facteur de prise en compte des enjeux éthiques, ou si ceux-ci découlent directement des réglementations antérieures.

34 Par ailleurs, les résultats présentés doivent également tenir compte du biais de sélection inhérent à la méthodologie que nous avons adoptée. En effet, la sélection de 500 profils proposés par LinkedIn à la suite de notre requête dans le moteur de recherche de la plateforme introduit de facto un biais non négligeable, notamment concernant la distribution géographique des profils analysés. Ceci pourrait ainsi expliquer que la région de Montpellier ait pu émerger à ce point dans les résultats du fait que nous avons réalisé notre enquête dans cette même région. Notre localisation géographique a peut-être été prise en compte par l'algorithme de LinkedIn afin de nous proposer des profils géographiquement plus proches, même si cela est difficilement mesurable.

35 Nous gardons à l'esprit qu'une démarche empirique est orientée de manière plus ou moins intentionnelle (hypothèses ou postulats de départ, échantillons, formulation des questions, interprétation des résultats et des comportements, etc.), et que les résultats statistiques qui en découlent ne doivent pas être considérés comme des vérités scientifiques absolues, malgré tout le soin que nous avons apporté à l'élaboration de notre protocole de recherche exploratoire. Cela signifie que la simple répétition d'un thème de réponses ne nous autorise pas pour autant à en déduire une règle générale. De plus, malgré la garantie d'anonymat, les entretiens ont également pu être préparés en amont afin de s'assurer que les propos tenus soient conformes à la loi et aux dispositions prises par leur entreprise, ce qui peut nuire à une certaine spontanéité et/ou liberté des propos, et donc à ne pas refléter exactement les opinions des différents interviewés.

36 Dix ans après la première apparition du terme big data, les entreprises cherchent aujourd'hui et par tous les moyens à comprendre de mieux en mieux leurs cibles en collectant des données qui leur permettent désormais de mieux

prédire leurs comportements de consommation. Et la technologie semble être sans limites. Alors que un français sur cinq renoncerait à l'utilisation d'outils numériques[7], ces 12 millions de nos concitoyens n'échappent pas pour autant au traitement de leurs données personnelles (paiements bancaires, sécurité sociale, programmes de fidélité, etc.), le plus souvent à leur insu. L'éthique de la collecte et du traitement algorithmique des données personnelles, au-delà du fait qu'elle soit désormais soumise à une réglementation européenne, semble être aujourd'hui une réelle préoccupation pour les experts des données massives que sont les data scientists au regard des résultats que nous avons obtenus. Enjeu à la fois éthique et économique, le secteur de la science de la donnée doit désormais parvenir à concilier une exploitation performante et sécurisée des données tout en respectant les contraintes inhérentes au cadre du RGPD.

Bibliographie

Babinet, G. (2014) L'ère numérique, un nouvel âge de l'humanité. Paris : Le Passeur.

Cardon D. (2012) « Regarder les données », Multitudes, n° 49, pp. 138–142.
DOI : 10.3917/mult.049.0138

Cecere G., Le Guel F., Rochelandet F. (2015), « Les modèles d'affaires numériques sont-ils trop indiscrets ? Une analyse empirique », Réseaux, n° 189, pp. 77-101.

Credoc (2017) « Baromètre du numérique 2018 ». Disponible en ligne sur : https://labo.societenumerique.gouv.fr/wp-content/uploads/2018/12/barometredunumerique2018.pdf (dernière consultation : 13 janvier 2019)

Dubois L., Gaullier F. (2017) « Publicité ciblée en ligne, protection des données à caractère personnel et ePrivacy : un ménage à trois délicat », Legicom, n° 59, pp. 69-102.
DOI : 10.3917/legi.059.0069

Jomni A. (2018) « Le Darknet est-il une zone de non droit ? », Sécurité globale, n° 15, pp. 17-23.
DOI : 10.3917/secug.183.0017

Khatchatourov A. (2019) Les identités numériques en tension. Entre autonomie et contrôle. Londres : ISTE éditions.

Le Moënne C. (2018) « Penser l'artificialisation du monde ? Retour sur la question des constructivismes et de la transformation numérique », Communication & Organisation, n ° 53, pp. 107-132.

Simonin P.M. (2015) « Information Technology Ethics : Concepts and Practices in a Digital World », Cambridge Scholars Publishing, Systèmes d'information & management, vol. 20, pp. 123-125.

Van Campenhoudt L., Marquet J., Quivy R. (2011) Manuel de recherche en sciences sociales. Paris : Dunod.

Verlaet L. (2015) « La deuxième révolution des systèmes d'information : vers le constructivisme numérique », Hermès, La Revue, n° 71, pp. 249-254.

Zolynski C. (2015) « Big data : pour une éthique des données », I2D – Information, données & documents, vol. 52, n° 2, pp. 25-26.

Notes

1 Computing Community Consortium (Bryant, Katz et Lazowska, 2008), « *Big-Data Computing : Creating Revolutionary Breakthroughs in Commerce, Science and Society* », disponible en ligne sur : http://www.cra.org/ccc/docs/init/Big_Data.pdf.

2 Rapport IDC pour Seagate (2018) « *The Digitization of the World From Edge to Core* » disponible en ligne sur : https ://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

3 Wall Street Journal (2019), « *You Give Apps Sensitive Personal Information. Then They Tell Facebook* », à lire en ligne sur : https ://www.wsj.com/articles/you-give-apps-

sensitive-personal-information-then-they-tell-facebook-11550851636

4 L'ICANN (« *Internet Corporation for Assigned Names and Numbers* ») se définit comme une organisation à but non lucratif et reconnue d'utilité publique rassemblant des participants du monde entier qui œuvrent à la préservation de la sécurité, la stabilité et l'interopérabilité de l'Internet : https ://www.icann.org/fr

5 Cnil (2018), « *RGPD : quel premier bilan 4 mois après son entrée en application ?* », consultable en ligne sur : https ://www.cnil.fr/fr/rgpd-quel-premier-bilan-4-mois-apres-son-entree-en-application

6 Classement « *Best jobs in America* », consultable en ligne sur https ://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

7 Baromètre du numérique 2018, rapport complet disponible en ligne sur : https://labo.societenumerique.gouv.fr/wp-content/uploads/2018/12/barometredunumerique2018.pdf

Pour citer cet article

Référence électronique
Nadia Hassani, « Le paradoxe de la protection des données personnelles à l'heure de la libre circulation des informations », *Terminal* [En ligne], 124 | 2019, mis en ligne le 30 juin 2019, consulté le 24 mars 2020. URL : http://journals.openedition.org/terminal/4040 ; DOI : https://doi.org/10.4000/terminal.4040

Auteur

Nadia Hassani
Laboratoire Lerass – Ceric, Sciences de l'Information et de la Communication, Université Paul Valéry Montpellier 3

1186 route de Mende

34199 Montpellier Cedex 5 - France

nadia.hassani1@gmail.com

Droits d'auteur