SAMMAN

Blaeu privacy response team

ÉTUDE · JUIN 2026

De la donnée pseudonymisée à la donnée anonyme

Un cadre d'évaluation pratique fondé sur l'arrêt CRU, appliqué à des cas d'usage réels fournis par des entreprises européennes

AUTEURS

Dr Rob van Eijk · Team Blaeu

Benjamin de Vanssay · SAMMAN Law & Corporate Affairs

SAMMAN Law & Corporate Affairs | cabinet-samman.com

SAMMAN Law & Corporate Affairs est un cabinet d'avocats basé à Paris et à Bruxelles, qui conseille ses clients en matière d'affaires publiques et réglementaires françaises et européennes. Le cabinet intervient à l'intersection du droit, des politiques publiques et de la stratégie institutionnelle, couvrant les dimensions législative, réglementaire et politique dans l'ensemble des secteurs.

Team Blaeu | blaeu.com

Fondé en 2019, Team Blaeu est un cabinet de conseil néerlandais spécialisé dans la protection des données de l'UE, la régulation des marchés numériques et la régulation de l'intelligence artificielle. Team Blaeu accompagne les organisations dans leur mise en conformité avec le RGPD, la gouvernance de l'*AI Act* et les procédures d'exécution aux niveaux national et européen.

Table des matières

Préambule 5

1. Introduction 6

1.1 La pseudonymisation sous l'approche absolue de la notion de donnée personnelle 6

1.2 La pseudonymisation sous l'approche relative consacrée par l'arrêt de la Cour de justice de l'Union européenne 8

1.3 La pseudonymisation en pratique et la nécessité de lignes directrices fondées sur des données probantes 8

2. De l'arrêt CRU et des lignes directrices des autorités de contrôle vers un cadre d'évaluation pratique 10

3. Cas d'usage 1 : Transmission de *prompts* expurgés à un fournisseur d'IA générative sous régime de non-conservation de données 13

3.1 Le fonctionnement de l'accord *Zero-Data-Retention (ZDR)* 13

3.2 La suppression des identifiants et la non-conservation font obstacle à l'individualisation et à la corrélation 14

3.3 Pourquoi les données d'entraînement ne peuvent constituer une voie de réidentification 15

3.4 Les *prompts* expurgés sous régime de non-conservation de données (*Zero Data Retention*, ZDR) 15

4. Cas d'usage 2 : Développement de cartes de trafic en temps réel à partir de signaux de véhicules anonymisés 17

4.1 La production de données par les véhicules flottants 18

4.2 Le regroupement aléatoire par lots et les intervalles dissimulent le début et la fin des trajets 19

4.3 La randomisation temporelle et la suppression du système central (*backend)* rompent la traçabilité des données 19

4.4 La conversion des points GPS en segments routiers supprime la localisation précise 20

4.5 L'agrégation collective (*swarm aggregation*) dissout les véhicules individuels dans les schémas de trafic 20

4.6 L'état des routes survit à la chaîne de traitement, les conducteurs individuels non 21

5. Cas d'usage 3 : Développement de fonctionnalités intelligentes de détection d'anomalies dans un logiciel métier hébergé dans le *cloud* à partir de données relationnelles 23

5.1 Les limites des technologies standard renforçant la protection de la vie privée (PET – *Privacy Enhancing Technologies*) 24

5.2 Le hachage salé et la séparation stricte des environnements 24

5.3 Les schémas atteignent l'entraînement, mais le responsable du traitement conserve la clé 24

6. Cas d'usage 4 : Transmission de statistiques d'audience agrégées à une plateforme éditoriale 26

6.1 La collecte et le traitement des données par le prestataire spécialisé dans la mesure d'audience 27

6.2 L'agrégation des données comme principale garantie de protection de la vie privée . 28

6.3 Quatre garde-fous convergent pour placer les rapports hors du champ d'application du RGPD pour la Plateforme 30

7. Cas d'usage 5 : Entraînement de modèles de prédiction publicitaire sans données des utilisateurs individuels 32

7.1 Le remplacement des identifiants directs par des identifiants hachés ou aléatoires ... 33

7.2 Le regroupement des enregistrements individuels en comptages agrégés supprime les données individuelles 34

7.3 L'ajout de bruit statistique protège la contribution de chaque personne 34

7.4 La reconstruction des signaux d'entraînement à partir des décomptes agrégés évite le recours à des données au niveau de l'utilisateur 35

7.5 Le jeu de données synthétiques au regard de l'évaluation en deux étapes..... 35

8. Conclusions. Ce que les cas d'usage révèlent sur les données pseudonymisées en pratique..... 37

8.1 Le dénominateur commun des cinq cas d'usage 38

8.2 Perspectives..... 38

GLOSSAIRE DES TERMES TECHNIQUES 40

ANNEXES..... 42

A : LE CONTENTIEUX CRU DEVANT LA CJUE 42

A1 : CONSEIL DE RÉSOLUTION UNIQUE (CRU) CONTRE CONTRÔLEUR EUROPÉEN DE LA PROTECTION DES DONNÉES 42

A2 : CONTRÔLEUR EUROPÉEN DE LA PROTECTION DES DONNÉES CONTRE CONSEIL DE RÉSOLUTION UNIQUE 42

A3 : CONCLUSIONS DE L'AVOCAT GÉNÉRAL..... 43

B : JURISPRUDENCE DE LA CJUE..... 43

B1 : SCARLET EXTENDED SA CONTRE SOCIÉTÉ BELGE DES AUTEURS, COMPOSITEURS ET ÉDITEURS SCRL (SABAM) 43

B2 : PATRICK BREYER CONTRE BUNDESREPUBLIK DEUTSCHLAND 43

B3 : PETER NOWAK CONTRE DATA PROTECTION COMMISSIONER 43

B4 : GESAMTVERBAND AUTOTEILE-HANDEL E.V. CONTRE SCANIA CV AB 44

B5 : OC CONTRE COMMISSION EUROPÉENNE..... 44

B6 : IAB EUROPE CONTRE GEGEVENSBESCHERMINGSAUTORITEIT 44

C : DÉFINITIONS ET CONSIDÉRANTS PERTINENTS DU RGPD..... 44

C1 : CONSIDÉRANT 26 DU RGPD 44
C2 : ARTICLE 4, POINT 1, DU RGPD – DONNÉES À CARACTÈRE PERSONNEL 45
C3 : ARTICLE 4, POINT 5, DU RGPD – PSEUDONYMISATION 45
C4 : ARTICLE 25 DU RGPD – PROTECTION DES DONNÉES DÈS LA CONCEPTION ET PROTECTION DES DONNEES PAR DÉFAUT 45

Préambule

L'arrêt CRU de la Cour de justice de l'Union européenne,[1] a relancé le débat de longue date sur les contours du règlement général sur la protection des données (RGPD). Il a confirmé que le caractère identifiable d'une donnée doit s'apprécier au regard de « l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que l'individualisation (*singling out*) ».[2] Cette appréciation est propre à chaque destinataire et donc contextuelle. Le présent rapport s'appuie sur cet arrêt et décline son application concrète en adoptant l'approche la plus restrictive lorsque les lignes directrices du CEPD et la jurisprudence divergent. La sécurité juridique exige des critères clairs pour déterminer à quel moment une donnée pseudonymisée devient une donnée anonyme. La présente étude propose une contribution pratique et fondée sur des données probantes à cette appréciation.

Elle s'appuie sur des cas d'usage fournis par des acteurs industriels européens de premier plan, dans les secteurs de l'automobile, du logiciel, du divertissement, de la legal tech et de la publicité, à l'issue d'échanges approfondis avec des experts techniques. Ces cas détaillent l'architecture technique utilisée par ces entreprises pour mettre en œuvre la pseudonymisation en pratique, et décrivent les couches de garanties qu'elles déploient contre la réidentification.

Pris ensemble, ces cas d'usage offrent un point de départ concret pour une application pratique et cohérente des exigences légales, au-delà des seules analyses de risque théoriques. En identifiant les schémas récurrents et les garanties efficaces issus de cas d'usage réels, l'étude jette les bases d'une évaluation plus structurée et plus prévisible des techniques d'anonymisation. Elle ouvre également la voie à l'élaboration de certifications et d'un cadre d'évaluation standardisé, fondé sur des critères clairs, opérationnels et vérifiables.[3]

Le présent rapport ne constitue pas un avis juridique. Ses auteurs sont des praticiens de la protection des données, non une autorité de contrôle. Les descriptions des cas d'usage reflètent l'état de la technologie au moment de la rédaction.

[1] Affaire C-413/23 P, Contrôleur européen de la protection des données contre Conseil de résolution unique, EU:C:2025:645.
[2] Considérant 26 du RGPD.
[3] Article 25, paragraphe 3, du RGPD, article 42 du RGPD.

1. Introduction

Près d'une décennie après son entrée en vigueur le 24 mai 2016 et son application depuis le 25 mai 2018, le règlement général sur la protection des données (RGPD) demeure l'une des initiatives réglementaires les plus ambitieuses et les plus influentes de l'ère numérique. En adoptant une approche fondée sur des principes, basée sur l'évaluation du risque et sur l'obligation de démontrer la conformité ("accountability"), le RGPD a vocation à adapter la directive 95/46 aux usages modernes de la donnée par les organisations, tout en préservant des garanties fortes pour les personnes et permettant la libre circulation des données au sein de l'Union européenne. Un large consensus se dégage parmi les parties prenantes quant à la pertinence des principes directeurs du RGPD et son architecture réglementaire face au rythme accéléré des cycles d'innovation.

Pourtant, malgré cette avance en matière réglementaire, l'Union européenne perd progressivement du terrain dans l''innovation fondée sur les données, en particulier dans le domaine de l'intelligence artificielle (IA). Comme cela a été souligné dans les débats politiques récents, y compris au plus haut niveau,[4] la compétitivité de l'Europe est contrainte par un environnement réglementaire complexe et fragmenté, générateur d'insécurité juridique.

Une grande partie de cette incertitude ne provient pas du RGPD lui-même, mais davantage de son interprétation rigide par les autorités de contrôle, qui se sont jusqu'ici abstenues de définir les contours de la notion centrale de donnée personnelle. La question est simple mais est sujet à controverse : dès lors que la notion de donnée non personnelle se définit négativement, c'est à dire comme toute donnée ne répondant pas à la qualification de donnée personnelle au sens du RGPD, que recouvre la notion de donnée non personnelle, et à partir de quand le RGPD cesse-t-il de s'appliquer ? Laisser cette question sans réponse paralyse l'application de l'ensemble des réglementations encourageant l'utilisation et le partage de données non personnelles, telles que le *Data Act* et le *Digital Markets Act*.

1.1 La pseudonymisation sous l'approche absolue de la notion de donnée personnelle

La notion de pseudonymisation occupe une position ambiguë en droit européen de la protection des données. Le droit de l'Union la reconnaît largement comme un instrument clé de réduction des risques et comme une garantie permettant le traitement sécurisé des données personnelles. Le RGPD la consacre à plusieurs reprises comme un moyen de réduire les risques, de renforcer la sécurité, et de mettre en œuvre des principes comme la protection des données dès la conception, la minimisation des données et la limitation des finalités. Plusieurs textes législatifs de l'Union font écho à cette reconnaissance.[5] Le RGPD

[4] Dans son rapport sur la compétitivité européenne, Mario Draghi identifie la complexité et la fragmentation réglementaires comme des freins structurels majeurs à la compétitivité de l'UE, générateurs d'insécurité juridique, d'une hausse des coûts de mise en conformité, et d'un effet dissuasif sur l'investissement dans les secteurs innovants. Mario Draghi, « The Future of European Competitiveness – A Competitiveness Strategy for Europe » (Commission européenne, 2024), https://commission.europa.eu/topics/competitiveness/draghi-report_en#paragraph_47059, consulté le 10 juin 2026.

[5] La pseudonymisation constitue actuellement un facteur clé permettant le traitement des données dans sept instruments juridiques majeurs de l'UE : le RGPD, le Data Act, le Data Governance Act et le Digital Services Act (partage et réutilisation responsables des données) ; le Digital Markets Act (interopérabilité des plateformes et échanges de données) ; la directive NIS2 (cybersécurité des systèmes d'information) ; et l'*AI Act* (entraînement des systèmes d'IA à haut risque).

n'exempte cependant les données pseudonymisées d'aucune des obligations du RGPD. Ce n'est que lorsque les données sont anonymisées que les obligations du RGPD cessent de s'appliquer. Mais l'anonymisation a malheureusement pour effet de priver les jeux de données de leur pertinence, laquelle est indispensable pour leur réutilisation ultérieure. En pratique, cependant, les autorités de contrôle n'ont pas clairement aidé les organisations à comprendre les différents critères permettant de distinguer les données pseudonymisées des données anonymes. Surtout, elles ont généralement traité les données comme étant personnelles dès lors qu'une réidentification potentielle d'une personne est théoriquement possible (approche absolue de la notion de donnée personnelle), même lorsque les responsables de traitement ne disposent d'aucun moyen juridique ou technique réaliste d'identifier les personnes et ont déployé des garanties robustes. Cette interprétation, fondée sur les lignes directrices des autorités de contrôle, s'est développée à rebours de l'approche fondée sur les risques que porte le RGPD.

Trois évolutions concomitantes aggravent cette incertitude. Premièrement, le Groupe de travail « Article 29 », prédécesseur de l'actuel Comité européen de la protection de données (CEPD), a publié un avis 05/2014 sur les techniques d'anonymisation (WP216, ci-après l'avis de 2014 sur l'anonymisation),[6] lequel est en attente depuis longtemps d'une mise à jour. Deuxièmement, le CEPD a finalisé ses lignes directrices 01/2025 sur la pseudonymisation. Troisièmement, les négociations en cours sur le *Digital Omnibus* pourraient aboutir à réviser la définition même de la donnée personnelle. Pris ensemble, ces développements fragilisent la capacité des entreprises à réutiliser leurs données à des fins d'innovation dans des conditions de sécurité juridique satisfaisantes.

Il en résulte que les organisations appliquent l'ensemble des obligations du RGPD quel que soit le risque réel, qu'il s'agisse de la tenue des registres des activités de traitement, de la réalisation des analyses d'impact relatives à la protection des données, de la négociation des contrats de traitements de données, du respect de la transparence ou des exigences de consentement. Cela alourdit inutilement les coûts de mise en conformité au détriment d'investissements en ingénierie ou en ressources techniques.

Cela freine en définitive la réutilisation des données et l'innovation, dans la mesure où la qualification d'une donnée comme donnée personnelle limite considérablement la capacité des responsables de traitement à réutiliser des données qu'ils détiennent déjà. Dans de nombreux cas, l'exemption relative à la recherche scientifique ne s'applique pas, car un traitement ultérieur réalisé à des fins de recherche peut très difficilement être présumé compatible avec la finalité initiale du traitement. Lorsque les données sont traitées sur le fondement de l'intérêt légitime, toute nouvelle finalité exige un test de mise en balance des intérêts. Le consentement constitue le cas le plus problématique car une nouvelle finalité exige l'obtention d'un nouveau consentement.

Cela pèse d'autant plus sur les entreprises européennes, en particulier les entreprises technologiques, qui disposent de ressources plus limitées que leurs concurrentes non européennes.

---

[6] Groupe de travail « Article 29 », avis 05/2014 sur les techniques d'anonymisation (WP216, 10 avril 2014). Le Groupe de travail réunissait les autorités de contrôle nationales en application de l'article 29 de la directive 95/46/CE ; le CEPD lui a succédé le 25 mai 2018 en application de l'article 68 du RGPD. Le CEPD a entériné un ensemble de lignes directrices du Groupe de travail dans sa décision d'entérinement 1/2018, mais le WP216 n'en faisait pas partie.

1.2 La pseudonymisation sous l'approche relative consacrée par l'arrêt de la Cour de justice de l'Union européenne

La Cour de justice de l'Union européenne (CJUE) a apporté un certain degré de clarification dans l'arrêt Conseil de résolution unique (CRU – SRB, *Single Resolution Board*).[7] L'arrêt a précisé que la notion de donnée personnelle comporte des limites claires. Les données pseudonymisées peuvent, dans certaines circonstances, faire effectivement obstacle à l'identification des personnes et être qualifiées d'anonymes. Dans cette affaire, la Cour a toutefois jugé que les données étaient personnelles du point de vue du CRU en tant que responsable du traitement, car celui-ci détenait l'ensemble des informations nécessaires pour réidentifier les personnes. La portée de l'arrêt réside dans le principe qu'il établit pour les destinataires de données pour lesquels les risques d'identification des personnes doivent être appréciés au cas par cas, laissant une large marge d'interprétation aux autorités de contrôle et aux juridictions.

Cette décision de la Cour représente une occasion unique de clarifier les notions de pseudonymisation et d'anonymisation et de garantir que les responsables de traitement puissent bénéficier de critères clairs leur permettant d'identifier quelles données relèvent du RGPD et lesquelles n'en relèvent pas. Elle permet de préserver en même temps un niveau élevé de protection des données personnelles tout en adoptant une approche plus pragmatique des obligations correspondantes, en excluant leur application dans les situations où un acteur ne dispose d'aucun moyen raisonnablement susceptible d'être utilisé pour identifier la personne.

En outre, la proposition de la Commission européenne visant à définir des critères harmonisés permettant de déterminer à quel moment des données pseudonymisées ne sont plus qualifiées de données personnelles, conformément à la jurisprudence de la CJUE, améliorerait, de la même manière, la sécurité juridique.[8] Elle garantirait également une meilleure harmonisation du RGPD en évitant des interprétations divergentes de l'arrêt CRU par les autorités de contrôle, les juridictions nationales et les autres autorités de régulation traitant de la pseudonymisation des données (autorités en charge de la concurrence, de la protection du consommateur, de l'information ou des télécommunications, par exemple).

1.3 La pseudonymisation en pratique et la nécessité de lignes directrices fondées sur des données probantes

La présente étude vise à étayer les arguments en faveur d'une anonymisation renforcée, y compris le développement des technologies renforçant la protection de la vie privée (PETs – *Privacy Enhancing Technologies*) comme outils complémentaires pour une réutilisation licite des données. L'étude est basée sur des exemples réels et concrets visant à fournir des orientations pratiques. L'objectif est de définir avec une plus grande précision les cas dans lesquels le RGPD s'applique et ceux dans lesquels il ne s'applique pas, afin d'alléger la charge de mise en conformité et de permettre aux acteurs européens de réutiliser leurs données avec davantage de confiance.

Les pratiques de surveillance nuisibles fondées sur des traitements de données intrusifs relèvent pleinement du champ d'application du RGPD ainsi que des autres dispositions

[7] Arrêt CRU (affaire C-413/23 P).
[8] Commission, « Digital Omnibus », COM(2025) 837 final, art. 41 bis. Article 41 bis proposé.

pertinentes du droit de l'Union et du droit national, indépendamment des procédés techniques qui les accompagnent. Le cadre (ci-après : Le Cadre d'Evaluation) présenté dans l'étude n'a pas vocation à requalifier de telles pratiques comme étant hors du champ d'application du RGPD. Elles appellent un renforcement de l'application des règles, non un allègement.

La pseudonymisation recouvre un large éventail de techniques, offrant chacune des niveaux différents de réduction du risque de réidentification. La mise en œuvre d'une seule technique de pseudonymisation n'est pas suffisante. Lorsque la donnée reste suffisante pour permettre l'individualisation d'une personne, les entreprises doivent adopter une approche à plusieurs niveaux. C'est précisément ce que démontre cette 'étude.

L'étude décrit des cas d'usage en détail, fournis par des praticiens, pour démontrer que le recours aux données pseudonymisées crée une valeur ajoutée tangible pour les organisations, stimule l'innovation et bénéficie à la société dans son ensemble, tout en permettant de documenter les risques résiduels. Ces cas d'usage devraient pouvoir s'appuyer sur la jurisprudence constante de la CJUE selon laquelle la donnée personnelle est une notion relative, et en particulier sur l'arrêt CRU qui précise quelles données pseudonymisées peuvent parfois sortir du champ d'application du RGPD. Notre objectif n'est pas de suggérer que la probabilité de réidentification est dénuée de tout fondement, mais d'encourager un débat politique ancré dans des éléments concrets et positifs plutôt que dans des affirmations théoriques et non étayées.

En s'appuyant sur des cas d'usage concrets, fournis par des entreprises européennes opérant dans différents secteurs avec une attention particulière au développement de l'IA et à la nécessité d'avoir accès aux données, la présente étude décrit comment la pseudonymisation opère en pratique : elle démontre comment elle réduit les risques tout en préservant l'utilité des données. Elle permet une réutilisation licite des données tout en respectant les standards de la CJUE sur l'anonymisation qui ne doit pas permettre l'identification des personnes pour sortir ainsi du champ d'application du RGPD.

La présente étude vise à aider les institutions concernées (la Commission européenne, le CEPD et toute autre partie prenante) à définir les critères harmonisés permettant de déterminer à quel moment des données pseudonymisées ne sont plus qualifiées de données personnelles.

La présente étude a été élaborée en concertation avec des entreprises européennes. Les cas d'usage ne sont pas purement théoriques. Chaque entreprise a apporté un cas d'usage concret. Chaque cas d'usage s'inscrit dans un contexte spécifique et constitue un point de référence clair pour toute analyse. Afin de permettre une évaluation cohérente entre les cas d'usage, nous présentons brièvement ci-dessous le cadre d'évaluation. Il fournit un ensemble de critères pratiques permettant d'évaluer la robustesse des mesures de pseudonymisation.

Nous présentons cinq cas d'usage. Ensemble, ils couvrent la legal tech, l'automobile, le logiciel d'entreprise (deux cas), la mesure d'audience et la technologie publicitaire. Le premier cas d'usage porte sur la transmission de *prompts* expurgés à un fournisseur d'IA générative sous régime de non-conservation des données (cas d'usage 1, section 3). Le deuxième cas d'usage porte sur le développement de cartes de trafic en temps réel à partir de signaux de véhicules anonymisés (cas d'usage 2, section 4). Le troisième porte sur le développement de fonctionnalités intelligentes de détection d'anomalies dans un logiciel métier dans le *cloud* à partir de données relationnelles (cas d'usage 3, section 5). Le quatrième porte sur la transmission de statistiques d'audience agrégées à une plateforme éditoriale (cas d'usage 4,

section 6). Le cinquième porte sur l'entraînement de modèles de prédiction publicitaire sans données individuelles des utilisateurs (cas d'usage 5, section 7).

2. De l'arrêt CRU et des lignes directrices des autorités de contrôle vers un cadre d'évaluation pratique

La pseudonymisation repose avant tout sur des mesures techniques et organisationnelles qui mettent en œuvre les principes de protection des données personnelles dès la conception et par défaut au titre de l'article 25, paragraphes 1 et 2 du RGPD. L'anonymisation, quant à elle, touche à la définition même de la donnée personnelle à l'article 4(1) du RGPD. Lorsque les données sont anonymisées et que la personne n'est pas ou n'est plus identifiable, le RGPD ne s'applique pas. Le seuil à atteindre pour rendre des données anonymes et les faire sortir du champ d'application du RGPD est donc bien plus élevé que celui de la pseudonymisation. Les deux notions, donnée personnelle et donnée anonyme, remplissent des fonctions distinctes et produisent des conséquences juridiques différentes, ce que le considérant 28 du RGPD confirme. Il dispose que « la pseudonymisation des données à caractère personnel peut réduire les risques pour les personnes concernées et aider les responsables du traitement et les sous-traitants à remplir leurs obligations en matière de protection des données. L'introduction explicite de la pseudonymisation dans le présent règlement ne vise pas à exclure toute autre mesure de protection des données ».

L'article 4(5) du RGPD décrit une relation, non une propriété fixe de la donnée. La 'pseudonymisation' désigne « 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 ».[9] La pseudonymisation présuppose que les informations supplémentaires nécessaires pour réidentifier les personnes existent (encore), qu'elles soient conservées séparément, et qu'elles fassent l'objet de mesures techniques et organisationnelles. Une donnée pseudonymisée peut donc constituer une donnée personnelle pour la partie qui détient ces informations supplémentaires, et une donnée anonyme pour un destinataire qui ne les détient pas et qui ne dispose d'aucun moyen raisonnablement susceptible de les obtenir. C'est là le cœur de l'arrêt CRU. La Cour a jugé que les données pseudonymisées ne sont pas, dans tous les cas et pour toute partie, des données personnelles, et que le caractère identifiable dépend des moyens dont dispose le destinataire concerné.[10] Notre cadre d'évaluation repose sur les critères définis par cet arrêt . Nous y ajoutons ensuite, par mesure de précaution, le seuil plus strict défini par l'avis de 2014 du CEPD sur l'anonymisation, pour les raisons exposées ci-dessous.

L'arrêt CRU détermine également le test d'identification applicable à un destinataire des données, et y attache deux conditions. La Cour a jugé que des données pseudonymisées transférées à un destinataire peuvent être non personnelles pour celui-ci. Premièrement, le destinataire ne doit pas être en mesure de lever les mesures pendant qu'il traite les données.

[9] Voir également l'annexe C3.
[10] Affaire C-413/23 P, EU:C:2025:645, points 75 à 77 et 86.

Deuxièmement, ces mesures doivent effectivement empêcher le destinataire de relier ces données à une personne physique, y compris par recoupement avec d'autres facteurs.[11] Les deux conditions sont cumulatives. Une mesure que le destinataire peut lui-même neutraliser, ou qui laisse subsister une voie d'identification par d'autres données, échoue au test d'anonymité. Nous appliquons ces deux conditions à chaque cas d'usage.

Nous évaluons l'anonymisation du point de vue du destinataire au regard de sept critères. Les six premiers reposent principalement sur l'avis de 2014 du CEPD sur l'anonymisation et interrogent le degré de difficulté qu'aurait, en pratique, une réidentification. Le septième critère repose sur l'arrêt CRU et interroge les moyens raisonnablement susceptibles d'être utilisés, par le destinataire ou par une autre personne, question à laquelle le considérant 26 du RGPD invite précisément à répondre.

1) Les jeux de données contenant des données pseudonymisées devraient être conservés séparément des données identifiant directement une personne physique.[12]
2) L'anonymisation devrait être irréversible pour le destinataire. Le cadre d'évaluation apprécie l'irréversibilité du point de vue du destinataire et non de façon abstraite. Lorsque le destinataire ne peut lever les mesures pendant son traitement et ne dispose d'aucun moyen raisonnablement susceptible de les inverser, les données sont irréversibles pour ce destinataire, même si le responsable du traitement conserve, pour ses propres finalités, les moyens de les inverser.[13]
3) La probabilité de réidentifier une personne physique devrait être insignifiante.[14]
4) La probabilité d'individualiser une personne physique devrait être faible. Une simple possibilité théorique d'individualisation n'est pas suffisante [15]
5) La probabilité de corréler des enregistrements relatifs à une personne physique, au sein d'un jeu de données ou entre plusieurs jeux de données, devrait être négligeable.[16]
6) La probabilité d'inférer, avec une probabilité significative, des informations sur une personne physique à partir des données devrait être négligeable.[17]

---

[11] Affaire C-413/23 P, point 77. La Cour a posé deux conditions cumulatives pour le destinataire. « S'agissant de Deloitte, à laquelle le CRU a transmis des commentaires pseudonymisés, les mesures techniques et organisationnelles visées à l'article 3, paragraphe 6, du règlement 2018/1725 peuvent, ainsi que le soutient pour l'essentiel le CRU, avoir pour effet que, pour cette société, ces commentaires ne revêtent pas un caractère personnel. Cela suppose toutefois, premièrement, que Deloitte ne soit pas en mesure de lever ces mesures lors d'un quelconque traitement des commentaires effectué sous son contrôle. Deuxièmement, ces mesures doivent effectivement être de nature à empêcher Deloitte d'attribuer ces commentaires à la personne concernée, y compris par le recours à d'autres moyens d'identification tels que le recoupement avec d'autres facteurs, de telle sorte que, pour cette société, la personne concernée ne soit pas ou plus identifiable. »

[12] Article 4(5) du RGPD.

[13] Voir l'avis de 2014 sur l'anonymisation. Voir également l'affaire C-413/23 P, point 77, sur l'irréversibilité appréciée du point de vue du destinataire.

[14] Voir l'arrêt CRU (affaire C-413/23 P), point 82 : « Par ailleurs, la Cour a déjà jugé qu'un moyen d'identifier la personne concernée n'est pas raisonnablement susceptible d'être utilisé lorsque le risque d'identification paraît en réalité insignifiant, en ce que l'identification de cette personne concernée est interdite par la loi ou impossible en pratique (...) ». Et affaire C-413/23 P, conclusions de l'avocat général Spielmann, point 57 : « Ainsi, ce n'est que lorsque le risque d'identification est inexistant ou insignifiant que des données peuvent légalement échapper à la qualification de 'données à caractère personnel'. »

[15] Le ciblage, la corrélation et l'inférence constituent les notions clés de l'avis de 2014 sur l'anonymisation.

[16] ibid

[17] ibid

7) L'évaluation devrait tenir compte des moyens raisonnablement susceptibles d'être utilisés, en pondérant les coûts d'identification, le temps qu'elle requiert, la technologie disponible au moment du traitement, ainsi que les évolutions technologiques, y compris la disponibilité de jeux de données exploitables par machine, de l'IA générative et de l'IA agentique.[18]

L'arrêt CRU et l'arrêt Breyer posent l'un comme l'autre la question de savoir si des moyens spécifiques de réidentification sont raisonnablement susceptibles d'être utilisés par un destinataire spécifique.[19] Dans l'arrêt Breyer, la Cour a jugé que les adresses IP dynamiques constituaient des données personnelles pour le responsable du traitement parce que celui-ci disposait de moyens *juridiques* lui permettant d'obtenir, auprès d'un tiers, des informations d'identification. Le raisonnement vaut dans les deux sens. Lorsque l'accès juridique à cette information fait clairement défaut, quand bien même l'information existerait quelque part, les données ne sont pas personnelles pour ce destinataire. Il convient de noter que la CJUE ne raisonne pas en seuils de probabilité, en scores de risque technique, ni en indicateurs de risque résiduel.

Une limite découle directement du considérant 26 du RGPD et du point 85 de l'arrêt CRU. L'évaluation est propre à chaque destinataire et le verdict s'attache au destinataire et non à la donnée elle-même. Une conclusion d'anonymat vaut pour un destinataire donné, dans un contexte donné. Elle ne survit pas à un transfert ultérieur vers une partie qui, elle, dispose de moyens raisonnablement susceptibles d'identifier la personne concernée. Le considérant 26 du RGPD vise les moyens dont dispose le responsable du traitement ou une autre personne, et l'arrêt CRU confirme que des données non personnelles pour un destinataire peuvent demeurer des données personnelles entre les mains d'un autre. Chaque nouveau destinataire doit mener sa propre évaluation.

L'avis de 2014 du CEPD sur l'anonymisation traite l'irréversibilité comme une exigence et pose l'individualisation, la corrélation et l'inférence comme des tests distincts. Il n'impose pas de méthode technique particulière, mais demande au responsable du traitement de satisfaire chaque test au regard des éléments de preuve. Nous ajoutons ce seuil plus strict au-dessus du test CRU, par mesure de *précaution*. En pratique, le cadre se déroule en deux étapes. Premièrement, nous appliquons le test CRU. Si le destinataire dispose de moyens raisonnablement susceptibles d'identifier la personne concernée, les données sont personnelles, l'évaluation s'arrête là. Deuxièmement, lorsque le test CRU oriente vers une conclusion d''anonymat, nous appliquons le seuil de l'avis de 2014 du CEPD sur l'anonymisation avant de confirmer ce résultat. Les données qui satisfont le test CRU mais échouent au seuil de l'avis de 2014 ne franchissent pas notre cadre d'évaluation. Nous les traitons comme des données très vraisemblablement non anonymes, car la précaution doit primer dans les cas limites même si cela signifie une réduction du champ d'application de l'anonymité.

Le cadre d'évaluation est construit pour pouvoir aboutir à un verdict négatif, et c'est le cas dans certains cas d'usage. Des mesures robustes ne suffisent pas à elles seules à passer le test. Dans certains cas, lorsque la réversibilité subsiste pour le responsable de traitement et que le destinataire peut potentiellement avoir accès à ces moyens de réversibilité, ou lorsque des données résiduelles dans un champ libre permettent potentiellement d'inférer des

[18] Article 4, point 5, du RGPD, affaire C-413/23 P, points 79 à 87, et le glossaire.
[19] Affaire C-582/14, Patrick Breyer contre Bundesrepublik Deutschland, EU:C:2016:779.

informations, le principe de précaution oriente la qualification des données vers des données personnelles et ce, malgré les mesures appliquées. Nous signalons ces cas comme très vraisemblablement non anonymes et en expliquons les raisons dans les sections concernées. Nous estimons que ces cas d'usage revêtent une importance particulière, car ils contribuent à l'objectif affiché de jeter les bases d'une évaluation plus structurée et plus prévisible des techniques d'anonymisation.

3. Cas d'usage 1 : Transmission de *prompts* expurgés à un fournisseur d'IA générative sous régime de non-conservation de données

Cadre d'évaluation

Évalués au regard du cadre exposé dans la Section 2, les *prompts* expurgés franchissent les deux étapes. Nous estimons qu'ils sont très vraisemblablement anonymes entre les mains du fournisseur d'IA générative en tant que destinataire, tout en demeurant des données personnelles pour le prestataire d'intelligence juridique qui détient les comptes utilisateurs.

Résumé. Ce cas d'usage a été fourni par un prestataire européen d'intelligence juridique, pour promouvoir de l'utilisation d'agents conversationnels IA en respectant la protection des données dès la conception. Lorsque des utilisateurs professionnels interagissent avec son agent conversationnel juridique, leurs prompts peuvent contenir des données personnelles (parfois sensibles). Le cas d'usage montre comment des garanties techniques permettent de supprimer les identifiants de ces prompts et d'assurer leur suppression immédiatement après traitement. Il en résulte que le fournisseur d'IA ne dispose d'aucun moyen réaliste de savoir qui a envoyé le prompt ni de le relier à une personne. Cela ne vaut que pour ce fournisseur et que pour les prompts, non pour les pièces jointes. Traiter de telles données comme non personnelles, lorsque ces garanties sont effectives, peut favoriser l'adoption en Europe d'outils d'IA respectueux de la vie privée dès la conception, en particulier pour les services qui s'appuient sur des modèles d'IA externes.

L'organisation est un prestataire d'intelligence juridique qui propose un service d'agent conversationnel juridique à des utilisateurs professionnels. Dans le cadre du service, l'organisation transmet les *prompts* générés par ses utilisateurs à des modèles d'IA générative de pointe (les modèles les plus performants disponibles). Avant l'envoi du *prompt*, l'organisation supprime tous les identifiants directs du contenu du *prompt*. Les *prompts* ne comportent ni nom, ni adresse électronique, ni identifiant utilisateur, ni aucun autre identifiant unique. Ils ne contiennent que la substance d'une question juridique ou factuelle.

Ce cas d'usage examine comment un service d'intelligence juridique agit en qualité de responsable du traitement pseudonymisant les données lorsqu'il transmet les *prompts* des utilisateurs à des modèles d'IA générative de pointe.

3.1 Le fonctionnement de l'accord *Zero-Data-Retention (ZDR)*

L'organisation applique un accord de non-conservation des données (*Zero-Data-Retention*, ZDR) à l'ensemble des fournisseurs d'IA générative. Les *prompts* sont traités en mémoire volatile (mémoire de travail temporaire qui disparaît à la fin du traitement) et ne sont ni stockés,

ni journalisés, ni indexés, ni conservés au-delà de la durée de l'inférence. L'organisation désactive également, contractuellement et techniquement, toute fonctionnalité côté fournisseur susceptible de permettre une réutilisation ou une valorisation du contenu des *prompts*, y compris les dispositifs de détection des abus, les processus d'amélioration des produits et l'entraînement de modèles.

Les conditions générales habituelles des plateformes d'IA générative indiquent clairement que les fournisseurs de modèles d'IA peuvent se réserver le droit d'analyser les *prompts* par exemple à des fins de sécurité des contenus, de journaliser les métadonnées de session à des fins de facturation, de conserver temporairement en cache des couples *prompt*-réponse pour améliorer l'expérience utilisateur et de collecter des données d'interaction à des fins d'entraînement. Le contenu est parfois transmis pour examen humain. L'accès au modèle d'IA représente davantage que le modèle lui-même. Le modèle s'inscrit dans une architecture d'IA fournie par la plateforme qui l'héberge.

L'accord ZDR bloque ces canaux. Il ne se limite pas à une interdiction contractuelle. Le fournisseur doit faire respecter ces restrictions par des contrôles techniques au niveau de l'infrastructure, de sorte qu'une mise à jour système ou un contrat renégocié ne puisse les neutraliser silencieusement. Un droit d'audit ou une attestation indépendante des contrôles de conservation du fournisseur permettrait à l'organisation de vérifier cette application effective, et nous considérons qu'une telle clause constitue une bonne pratique pour tout accord ZDR.

3.2 La suppression des identifiants et la non-conservation font obstacle à l'individualisation et à la corrélation

Les *prompts* reçus par le fournisseur d'IA générative ne contiennent aucun identifiant direct et ne sont pas conservés au-delà du processus d'inférence dans le modèle d'IA de pointe. Cette combinaison fait structurellement obstacle à deux des trois vecteurs centraux de réidentification (l'individualisation et la corrélation). Le troisième (l'inférence, ou la réidentification par les données d'entraînement du modèle) est traité en section 3.3.

L'individualisation exige d'un adversaire qu'il combine des quasi-identifiants,[20] pour réduire un jeu de données à une seule personne. Les *prompts* transmis par l'organisation ne comportent aucun identifiant et portent, par construction, sur une matière juridique ou factuelle plutôt que sur des informations personnelles concernant l'utilisateur du service. Aucun quasi-identifiant n'est disponible à partir duquel individualiser une personne physique.

Une question légitime se pose alors. Que se passe-t-il si la même personne envoie des *prompts* similaires à d'autres services qui, eux, conservent des identifiants ? Quelqu'un pourrait alors tenter de faire correspondre ces *prompts* nominatifs aux *prompts* expurgés, en s'appuyant sur les termes employés dans chaque *prompt* et sur l'heure de leur envoi. Présentés ainsi, le contenu et le moment de l'envoi commencent à fonctionner comme des quasi-identifiants, même si le *prompt* lui-même ne porte ni nom ni compte.

Nous estimons que cette mise en correspondance peut malgré tout échouer pour ce fournisseur. Dès lors qu'il ne détient aucun de ces journaux externes et n'a aucun moyen licite

[20] Voir le glossaire.

de se les procurer, et qu'il ne conserve rien de son côté, il n'existe aucun *prompt* stocké auquel comparer quoi que ce soit. Avec un seul côté du couple, il n'y a rien à recouper.

La corrélation exige l'accès à des enregistrements stockés pouvant être recoupés avec des jeux de données externes. La politique ZDR élimine la couche d'accumulation de données dont dépend structurellement toute attaque par corrélation. En l'absence de données persistantes, il n'y a rien à corréler.

3.3 Pourquoi les données d'entraînement ne peuvent constituer une voie de réidentification

Les techniques de réidentification reposent souvent sur l'accumulation de données, le recoupement et la reconnaissance de schémas au sein de jeux de données stockés. En l'absence de toute donnée persistante, ces techniques ne peuvent être mises en œuvre. Il subsiste un vecteur d'attaque supplémentaire à considérer : le *prompt* existe encore pendant son trajet de l'organisation vers le fournisseur. Un attaquant présent sur le réseau pourrait tenter de l'intercepter durant ce transfert. Pour l'empêcher, l'organisation devrait envoyer chaque *prompt* via un canal chiffré, afin que nul ne puisse lire la donnée en transit. Il s'agit d'une mesure de sécurité, et non d'une question relative au caractère identifiable, mais elle couvre le seul moment que la politique ZDR ne couvre pas.

Une autre question se pose du fait que les fournisseurs d'IA générative exploitent des modèles entraînés sur de vastes jeux de données, susceptibles d'inclure du contenu juridique publiquement disponible. La question se pose alors de savoir si de telles données d'entraînement pourraient constituer un vecteur de réidentification (par exemple, en corrélant la substance ou le style d'une requête avec des personnes identifiables dont les écrits ou les affaires figurent dans le corpus d'entraînement). Techniquement, les grands modèles de langage ne fonctionnent pas comme une recherche dans une base de données restituant un enregistrement stocké. La véritable question est de savoir si un attaquant pourrait reconstituer un identifiant supprimé à partir des mots qui l'entourent. Des travaux de recherche récents montrent que cela pourrait être possible.[21] Sur le plan juridique, tout usage délibéré des données d'entraînement à des fins de réidentification sortirait entièrement du cadre contractuel convenu avec la plateforme de l'organisation.

3.4 Les *prompts* expurgés sous régime de non-conservation de données (*Zero Data Retention*, ZDR)

Nous appliquons le cadre d'évaluation en deux étapes. Nous exécutons d'abord le test CRU, puis le seuil de précaution de l'avis de 2014 du CEPD sur l'anonymisation. Si le cas franchit les deux étapes, les données sont très vraisemblablement anonymes.

Étape 1, le test CRU : L'arrêt CRU pose deux conditions cumulatives pour le destinataire, ici le fournisseur d'IA générative. La première interroge la capacité du fournisseur à lever les

[21] Lucas Georges Gabriel Charpentier et Pierre Lison, « Re-identification of De-identified Documents with Autoregressive Infilling » (2025), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p. 1192. Cette étude a permis de reconstituer jusqu'à 80 % des éléments masqués dans des décisions de justice et d'autres documents, en comparant le texte restant à une vaste base de connaissances externes. Ce fournisseur ne dispose d'aucune base de ce type, et ne conserve aucun *prompt*, de sorte que cette attaque n'a rien sur quoi s'appuyer.

mesures pendant son traitement. Nous estimons qu'il ne le peut pas. L'organisation supprime tous les identifiants directs avant que le *prompt* ne quitte ses systèmes et l'accord ZDR impose au fournisseur de faire respecter la non-conservation par des contrôles techniques au niveau de l'infrastructure, et non par une seule clause contractuelle. L'organisation ne peut inspecter l'infrastructure du fournisseur, de sorte que cette partie de l'évaluation repose sur le respect, par le fournisseur, de cette obligation. Le fournisseur ne reçoit jamais l'information identifiante, et ne peut donc restituer ce qu'il n'a jamais détenu. La seconde condition interroge si les mesures empêchent effectivement le fournisseur d'attribuer le *prompt* à une personne physique, y compris par recoupement avec d'autres facteurs. Nous estimons qu'elles le font. Le *prompt* ne comporte ni nom, ni adresse électronique, ni identifiant utilisateur et le fournisseur ne dispose d'aucun jeu de données accumulé permettant un tel recoupement. Réidentifier un utilisateur à partir d'un *prompt* isolé, sans identifiant, traité sans conservation, exigerait un effort de calcul et d'investigation disproportionné. Les deux conditions sont réunies, de sorte qu'à notre avis le test CRU oriente vers l'anonymat pour ce destinataire.

Étape 2, le seuil de précaution de l'avis de 2014 sur l'anonymisation. Franchir le test CRU ne met pas fin à notre évaluation. Nous soumettons ensuite les données aux critères plus stricts de l'avis de 2014 avant d'accepter le résultat.

Cas d'usage 1 – Cadre d'évaluation CRU	
Critère	**Evaluation**
Séparation	L'organisation supprime tous les identifiants directs avant de partager le *prompt*. Le fournisseur ne reçoit que le contenu du *prompt* et ne détient aucune des données d'identité conservées par l'organisation. La séparation est structurelle, non simplement organisationnelle.
Irréversibilité	Le fournisseur traite les *prompts* en mémoire volatile et les supprime à la fin de l'inférence. Aucun enregistrement persistant ne subsiste à aucun stade de la chaîne. Rien ne pouvant être inversé, c'est l'architecture elle-même qui assure l'irréversibilité, et cela vaut pour le fournisseur lui-même plutôt que de reposer sur une clé conservée ailleurs.
Individualisation	L'individualisation exige un jeu de données à restreindre. La politique de non-conservation signifie qu'aucun *prompt* ne subsiste au-delà de la durée de l'inférence. Même un *prompt* au contenu factuel détaillé sur une affaire précise ne peut permettre de cibler son auteur, car le fournisseur ne conserve aucun jeu de données dans lequel mener une telle opération.
Corrélation	Aucun enregistrement ne subsiste au-delà de l'inférence. En l'absence d'enregistrements stockés, une corrélation entre jeux de données n'est pas possible en pratique.
Inférence	Les grands modèles de langage ne fonctionnent pas comme des systèmes de récupération d'information, et nul ne peut rétro-concevoir leurs poids statistiques pour attribuer un *prompt* à une personne

	précise. Le risque connexe d'attaque par inférence d'appartenance,[22] qui interroge si un *prompt* a fait partie des données d'entraînement, échoue également ici. L'accord interdit au fournisseur, contractuellement et techniquement, d'utiliser les *prompts* à des fins d'entraînement ou d'amélioration des modèles. Ces *prompts* n'entrent jamais dans le corpus d'entraînement, de sorte qu'il n'existe aucune appartenance à inférer.

Celles-ci confirment les deux étapes. Nous estimons donc que les *prompts* expurgés sont très vraisemblablement anonymes entre les mains du fournisseur d'IA générative, au sens du considérant 26 du RGPD.

Il convient toutefois de garder à l'esprit trois limites. Premièrement, l'évaluation ne porte que sur le traitement fondé sur les *prompts*. Elle exclut les pièces jointes, qui peuvent comporter un contenu de nature différente et appeler leur propre évaluation du caractère identifiable. Un document scanné ou une photo passée par reconnaissance optique de caractères (*Optical Character Recognition*), un outil qui transforme des images de texte en texte exploitable par machine, pourrait contenir des données personnelles que la suppression des identifiants des *prompts* ne voit jamais. Deuxièmement, le verdict est propre au destinataire. Il vaut pour ce fournisseur dans le cadre de cet accord et ne s'applique pas à un destinataire qui disposerait de moyens raisonnablement susceptibles d'identifier l'auteur. Troisièmement, l'expurgation ne supprime que les identifiants directs. La substance d'une question peut encore comporter des éléments permettant de remonter à une personne. Ce fournisseur ne peut agir sur ces éléments, car aucun *prompt* ne subsiste après l'inférence et il ne dispose de rien auquel comparer un *prompt*.

4. Cas d'usage 2 : Développement de cartes de trafic en temps réel à partir de signaux de véhicules anonymisés

Cadre d'évaluation

Au regard du cadre d'évaluation décrit dans la Section 2, les données de véhicules flottants (*floating car data*) franchissent les deux étapes entre les mains du prestataire tiers en tant que destinataire, dès lors que le seuil minimal de contributeurs est respecté. Les données issues de routes peu fréquentées qui n'atteignent pas ce seuil peuvent demeurer des données personnelles.

Résumé. Ce cas a été fourni par un constructeur automobile européen, à l'appui du développement de services de mobilité plus sûrs et plus intelligents. Les véhicules connectés peuvent générer des données sur l'état des routes, le trafic et les risques de sécurité, comme le verglas, données qui peuvent initialement comporter des signaux de localisation et liés au véhicule. Le cas d'usage montre comment des pseudonymes de courte durée, un regroupement aléatoire en lots, des délais temporels, la suppression des données brutes et l'agrégation entre de nombreux véhicules peuvent empêcher le prestataire tiers de relier les données à un véhicule ou un conducteur précis. Il en résulte que le prestataire reçoit des informations sur des segments routiers et des schémas de

[22] Voir le glossaire.

trafic, et non sur des trajets individuels. Traiter de telles données comme non personnelles, lorsque ces garanties sont effectives, peut favoriser le développement en Europe des services de mobilité plus sûrs et plus intelligents, tout en garantissant que le système voit la route, et non le conducteur. Sur les routes peu fréquentées comportant peu de véhicules, un seuil minimal de contributeurs supprime le point de donnée et les données qui n'atteignent pas ce seuil peuvent demeurer personnelles.

Ce cas d'usage examine si les données de véhicules flottants constituent des données personnelles entre les mains d'un prestataire tiers. Les données de véhicules flottants constituent une notion établie dans les politiques publiques et l'industrie européennes. Elles couvrent les données de capteurs que des véhicules en mouvement collectent et versent dans un fonds commun. Cette notion intervient dans les discussions relatives à l'Espace européen des données de mobilité.

Par exemple, un conducteur approchant une plaque de verglas ne peut la voir avant qu'il ne soit trop tard. Des millions de véhicules connectés comblent cette lacune. Ils partagent en continu des données d'adhérence et de performance au sein de vastes flottes, contribuant ainsi à la détection en temps réel des dangers sur la route et à la maintenance prédictive. Ces mêmes véhicules cartographient l'état des chaussées et les conditions de circulation au fil de leurs trajets, alimentant des services qu'une infrastructure routière fixe ne pourrait jamais couvrir à la même échelle.

Avant l'envoi des données, le système supprime les identifiants directs et les remplace par des pseudonymes de courte durée. Les données sont agrégées entre plusieurs véhicules par point de donnée. La randomisation temporelle et la suppression rapide des données brutes font le reste. À grande échelle, les observations qui se chevauchent créent un nuage de données dense au sein duquel les véhicules individuels deviennent indistinguables, rendant toute attribution pratiquement irréalisable.

4.1 La production de données par les véhicules flottants

Chaque véhicule de la flotte capte des données en continu pendant qu'il circule. Des caméras et des capteurs embarqués enregistrent l'horodatage, les positions GPS et l'état de la chaussée, tels que la qualité de la route et l'adhérence. Le véhicule transmet ces données au système central du constructeur en quelques secondes.

Les données utilisent un identifiant de véhicule haché (un hachage du numéro d'identification du véhicule - NIV). Le hachage est un procédé mathématique consistant à transformer une information en une chaîne de longueur fixe de lettres et de chiffres, difficilement réversible.[23] Le NIV est un code unique de 17 caractères qui identifie chaque véhicule dès sa sortie de chaîne de production. L'équivalent haché d'un NIV, par exemple « WVWZZZ1JZXW000001 », est la chaîne « 38657c765e00f5c9f487540df0aa7b36e08a0ec6db94de4658cb47eacca32ff3 ». Le système central du constructeur reçoit des données comportant cet identifiant haché et peut relier des paquets provenant d'un même véhicule pendant 30 minutes au maximum. Une

[23] Un NIV haché résiste à un examen superficiel, mais un NIV constitue une entrée courte, structurée et énumérable, de sorte qu'un hachage non salé reste vulnérable à une recherche par force brute sur cet espace d'entrée limité. Cela concerne le constructeur, qui détient l'identifiant haché pendant sa fenêtre de 30 minutes. Cela ne modifie pas la position du prestataire tiers, qui ne reçoit jamais le NIV haché et n'a donc rien sur quoi exercer une attaque par force brute.

consigne de travail contractuellement contraignante encadre ce que le personnel peut faire de ces données pendant cette fenêtre, ajoutant un contrôle organisationnel au contrôle technique. Une fois cette fenêtre close, le constructeur supprime l'identifiant haché des données. Aucun lien ne subsiste alors entre la donnée et le véhicule d'origine.

Une autre garantie de protection de la vie privée mérite d'être mentionnée : les conducteurs individuels peuvent désactiver le partage de données depuis l'intérieur du véhicule, ce qui leur donne une maîtrise directe de leur contribution.

4.2 Le regroupement aléatoire par lots et les intervalles dissimulent le début et la fin des trajets

Le véhicule ne transmet pas les données en continu. Le système central du constructeur instruit le système embarqué de regrouper les données par lots, avec des intervalles aléatoires délibérés entre les lots.

La randomisation fonctionne comme suit : la vitesse détermine les plages. Aux vitesses urbaines plus faibles, les lots s'étendent entre deux et cinq kilomètres, avec des intervalles moyens d'environ cinq cents mètres. Aux vitesses autoroutières supérieures à, par exemple, 110 kilomètres par heure, les lots s'étendent entre huit et douze kilomètres, avec des intervalles supérieurs à un kilomètre et demi. Le système ajoute en sus une variation aléatoire, de sorte qu'aucun schéma fixe n'apparaisse.

Cette conception protège l'origine et la destination de chaque trajet contre tout tiers ou acteur malveillant observant le flux de données. Le véhicule ne transmet rien durant les cinq cents premiers mètres d'un trajet, et n'émet aucun signal marquant la fin d'un trajet. Si un lot est incomplet en raison du processus de randomisation, il n'est pas transmis. Seuls des lots complets et pleinement constitués quittent le véhicule.

Deux termes reviennent tout au long de ce cas d'usage et portent des sens différents. Un *lot* désigne la fenêtre de collecte, le tronçon de route sur lequel le véhicule rassemble des données avant qu'un intervalle ne débute. Un *conteneur* désigne le paquet de transmission. Une fois un lot complété, le système enveloppe les données dans un conteneur, y joint le NIV haché, et l'envoie au système central du constructeur. Les deux termes décrivent la même unité de donnée à des étapes différentes.

4.3 La randomisation temporelle et la suppression du système central (*backend)* rompent la traçabilité des données

Le constructeur applique deux mesures avant de transmettre tout conteneur au prestataire tiers.

Premièrement, il applique un délai temporel aléatoire à chaque conteneur, de manière indépendante. Les conteneurs peuvent parvenir au système du prestataire tiers dans un ordre différent de celui du véhicule qui les a produits. Inverser ce réordonnancement exigerait un effort de calcul considérable.

Deuxièmement, une fois chaque conteneur transmis, le constructeur le supprime de son propre système central. Le constructeur n'en conserve aucune copie.

Avant qu'un conteneur ne quitte ses systèmes, l'identifiant de véhicule haché en a déjà été retiré. Le constructeur mesure la rapidité avec laquelle cette suppression intervient à l'échelle

de la flotte. À titre d'exemple, la moitié des données perd son identifiant de véhicule en un peu plus de deux heures, les trois quarts en huit heures et demie, neuf sur dix en moins de deux jours, et quatre-vingt-quinze pour cent en six jours. Le prestataire tiers ne reçoit que des données dépouillées, sans référence à un NIV.

Un adversaire qui parviendrait à déjouer le réordonnancement n'aurait toujours aucun moyen de dire quel véhicule a produit la séquence. Rien ne la rattache à un véhicule.

4.4 La conversion des points GPS en segments routiers supprime la localisation précise

Le prestataire tiers décompose chaque conteneur, fait correspondre ses coordonnées GPS au segment routier le plus proche sur une carte numérique, puis supprime le conteneur brut. Transformer une donnée véhicule en segment routier de cette manière constitue une étape déterminante. Le prestataire ne conserve aucune donnée brute. Le résultat prend la forme d'une couche cartographique ou d'un signal de sécurité reflétant les observations agrégées de nombreux véhicules sur un segment routier, et non l'historique de déplacement d'un véhicule en particulier.

Le prestataire tiers peut également agréger les données de segments routiers qu'il reçoit avec des données de véhicules flottants provenant d'autres constructeurs automobiles afin d'améliorer la couverture cartographique. Cela dissout davantage encore la contribution de tout véhicule individuel dans un ensemble plus large alimenté par plusieurs constructeurs automobile.

4.5 L'agrégation collective (*swarm aggregation*) dissout les véhicules individuels dans les schémas de trafic

L'agrégation collective[24] repose sur la déclaration simultanée d'observations par de nombreux véhicules pour les mêmes segments routiers. La flotte génère des observations qui se chevauchent, provenant de différents véhicules, sur chaque point du réseau routier. À cette échelle, aucun véhicule pris isolément ne produit un point de donnée qui se distingue des autres.

Des rapports qui se chevauchent, provenant de nombreux véhicules, créent un nuage de données. La contribution d'un véhicule individuel n'est qu'une contribution parmi d'autres pour un lieu donné. Même une tentative ciblée d'attribution se heurterait aux limites aléatoires des lots, aux intervalles aléatoires, au réordonnancement par délai temporel et à la simple densité des contributions des autres véhicules.

Un scénario mérite d'être abordé directement. Sur des routes peu fréquentées à faible trafic, telles que des routes rurales aux heures creuses, l'effet de l'agrégation collective s'affaiblit. Peu de véhicules contribuent des données pour le même segment et l'observation d'un seul véhicule peut ne pas se dissoudre dans un groupe plus large.

L'IA pourrait déceler des schémas de comportement de véhicules individuels que des analystes humains ne percevraient pas. L'architecture traite directement cette hypothèse. Le prestataire tiers applique un seuil minimal de contributeurs au niveau de la sortie, ne

[24] Voir le glossaire.

conservant une observation de segment routier que lorsqu'un nombre déterminé de véhicules y ont contribué dans une fenêtre temporelle donnée. En deçà de ce seuil, le prestataire supprime le point de donnée plutôt que de le conserver. Le cas limite d'un seul véhicule sur une route rurale calme, à une heure creuse, ne laisse aucune trace dans le jeu de données. Combinée à la suppression immédiate des conteneurs, aucune observation individuelle ne subsiste, même lorsque l'effet d'agrégation collective est ténu.

4.6 L'état des routes survit à la chaîne de traitement, les conducteurs individuels non

Nous exécutons d'abord le test CRU, puis le seuil de précaution de l'avis de 2014 du CEPD sur l'anonymisation. Le cadre d'évaluation se place du point de vue du prestataire tiers en tant que destinataire, car le constructeur automobile détient l'identifiant haché pendant sa fenêtre de 30 minutes et les données demeurent personnelles entre ses mains.

Étape 1. L'arrêt CRU pose deux conditions cumulatives pour le destinataire, ici le prestataire tiers. La première interroge la capacité du prestataire à lever les mesures pendant son traitement. Nous estimons qu'il ne le peut pas. Le constructeur retire l'identifiant de véhicule haché de chaque conteneur avant de le transmettre, de sorte que le prestataire ne reçoit jamais de NIV et ne peut restituer ce qu'il n'a jamais détenu. Le prestataire ne peut pas non plus annuler le regroupement aléatoire en lots, la suppression des limites de trajet, ni le réordonnancement par délai temporel, car le système embarqué et le système central du constructeur appliquent ces mesures en amont. La seconde condition interroge si les mesures empêchent effectivement le prestataire d'attribuer un conteneur à une personne physique, y compris par recoupement avec d'autres facteurs. Sur les routes fréquentées, c'est le cas. Les conteneurs comportent des coordonnées GPS et des horodatages, mais aucun identifiant ; les cinq cents premiers mètres de chaque trajet ne quittent jamais le véhicule ; aucun signal ne marque la fin d'un trajet ; et le prestataire associe chaque conteneur à une carte avant de supprimer la donnée brute. Il ne subsiste aucune trace persistante à recouper avec un jeu de données de localisation externe. Cette suppression relève de la propre pratique du prestataire, et ne pèse donc pas dans l'étape 1. Nous la comptabilisons comme un élément de l'étape 2, et les mesures en amont suffisent à elles seules à satisfaire les deux conditions sur les routes fréquentées.

Un scénario appelle une attention particulière. Sur une route rurale peu fréquentée, à une heure creuse, l'observation d'un seul véhicule peut se distinguer, et un schéma unique peut rouvrir une voie d'attribution. Le seuil minimal de contributeurs répond directement à cette hypothèse. Le prestataire ne conserve une observation de segment routier que lorsqu'un nombre déterminé de véhicules y ont contribué dans la fenêtre temporelle, et supprime le point de donnée en deçà de ce seuil. Ce seuil relève du propre contrôle du prestataire, et nous le comptabilisons donc comme un élément de l'étape 2, non comme une mesure de l'étape 1. Lorsqu'il s'applique, le risque lié aux routes peu fréquentées est neutralisé et le verdict de très vraisemblablement anonyme vaut pour ce destinataire. Lorsqu'il ne s'applique pas, les données issues de routes peu fréquentées peuvent demeurer des données personnelles, et nous estimons qu'elles devraient être traitées comme telles.

Étape 2. Nous soumettons ensuite les données aux critères de l'avis de 2014 sur l'anonymisation.

Cas d'usage 2 – Cadre d'évaluation CRU	
Critère	**Evaluation**
Séparation	Le prestataire ne reçoit jamais d'identifiant de véhicule. Le constructeur retire le NIV haché de chaque conteneur de données brutes avant de le transmettre, et le délai temporel aléatoire brouille l'ordre d'arrivée, de sorte que toute tentative de reconstituer la séquence d'un véhicule exigerait un effort de calcul considérable et constituerait également une violation du contrat. La séparation est structurelle, non simplement organisationnelle.
Irréversibilité	Aucun NIV n'atteint le prestataire, qui supprime chaque conteneur de données brutes une fois les points GPS associés à un segment routier. Rien ne subsiste à inverser, et cela vaut pour le prestataire lui-même, plutôt que sur une clé conservée ailleurs. Le constructeur conserve, lui, les moyens de relier des paquets pendant 30 minutes au maximum, mais cette capacité reste entre les mains du responsable du traitement, non du destinataire.
Réidentification	La probabilité de réidentifier une personne physique demeure très vraisemblablement insignifiante. Une tentative de réidentification devrait simultanément déjouer le regroupement aléatoire en lots, le réordonnancement des données et la densité de l'agrégation collective. En outre, la suppression des conteneurs de données brutes élimine tout point d'attaque résiduel.
L'individualisation	Les conteneurs ne comportent aucun quasi-identifiant. Il n'existe ni nom, ni identifiant d'appareil persistant, ni point de départ ou d'arrivée de trajet permettant de rattacher un enregistrement à un véhicule. Les longueurs de lots aléatoires et la suppression des limites de trajet n'offrent aucune prise à la réidentification, la détection de schémas par l'IA se heurte au même obstacle structurel.
Corrélation	Aucun pseudonyme persistant ne relie les observations entre conteneurs et le prestataire ne dispose d'aucune trace GPS brute à comparer avec des jeux de données externes. Chaque sortie par segment routier reflète la contribution combinée de nombreux véhicules et le prestataire peut en outre la mettre en commun avec des données de véhicules flottants provenant d'autres constructeurs, ce qui dissout davantage encore la part de tout véhicule pris isolément.
Inférence	Les données décrivent des segments routiers, non des conducteurs. La sortie agrégée ne dit rien de l'endroit où s'est rendu un véhicule particulier, de sa vitesse, ni de qui se trouvait au volant.

Celles-ci confirment les deux étapes lorsque le seuil minimal de contributeurs s'applique. Nous estimons que les données de véhicules flottants sont très vraisemblablement anonymes entre les mains du prestataire tiers en tant que destinataire, au sens du considérant 26 du RGPD, et qu'elles demeurent personnelles entre les mains du constructeur qui détient encore l'identifiant haché.

Deux précisions importantes encadrent cette conclusion. Premièrement, elle dépend du seuil minimal de contributeurs. La présente étude décrit un cas d'usage générique et ne fixe donc ni valeur de seuil ni fenêtre temporelle. Ces paramètres diffèrent d'un système à l'autre. Le verdict suppose donc que le prestataire fixe les deux à un niveau qui supprime effectivement les observations d'un seul véhicule et peut le démontrer sur demande. Sur une route rurale peu fréquentée à une heure creuse, en l'absence de cette suppression, le schéma d'un seul véhicule peut subsister, et les données peuvent demeurer personnelles. Deuxièmement, le verdict est propre au destinataire. Il vaut pour ce prestataire dans cette chaîne de traitement, et ne s'applique pas à un destinataire qui disposerait de moyens raisonnablement susceptibles d'identifier le véhicule ou son conducteur.

5. Cas d'usage 3 : Développement de fonctionnalités intelligentes de détection d'anomalies dans un logiciel métier hébergé dans le *cloud* à partir de données relationnelles

Cadre d'évaluation

Ce cas d'usage ne satisfait pas les critères du cadre d'évaluation. Il suffirait pourtant d'un seul ajustement pour les satisfaire pleinement, à savoir confier les clés à un tiers. Nous avons délibérément choisi de l'inclure pour montrer que le seuil du test est très élevé, mais non hors d'atteinte en pratique.

Résumé du cas. Ce cas a été fourni par un éditeur européen de logiciel métier hébergé dans le cloud, à l'appui du développement de fonctionnalités d'IA respectueuses de la vie privée. Les logiciels métiers intègrent de plus en plus des fonctionnalités d'IA qui permettent de détecter des schémas inhabituels dans les données clients et les signaler. Pour construire ces fonctionnalités, les modèles d'IA doivent s'entraîner sur des données de métier réelles. Ces données contiennent initialement des informations permettant d'identifier les utilisateurs individuels. Le cas d'usage montre comment le prestataire brouille les identifiants clients avant que les données n'atteignent l'équipe de développement de l'IA. Il verrouille la clé qui permettrait de reconstituer les données initiales dans un système distinct, de sorte que les développeurs d'IA ne disposent d'aucun moyen pratique de réidentifier qui que ce soit. L'équipe de développement extrait des enseignements des schémas contenus dans les données sans jamais savoir de qui il s'agit. Ce cas montre que même des garanties solides échouent dès lors que le responsable du traitement conserve la clé, de sorte que les données demeurent personnelles entre ses mains.

Ce cas d'usage examine comment un éditeur de logiciel métier hébergé dans le *cloud* traite des données clients relationnelles afin de développer des fonctionnalités intelligentes, telles que la détection d'anomalies, tout en testant les limites de l'anonymisation.

Des fonctionnalités intelligentes qui détectent les écarts par rapport à un comportement « normal » et qui proposent des actions correctrices font partie désormais de tout logiciel d'entreprise sur le marché. Pour construire de telles fonctionnalités, les modèles d'IA doivent être entraînés sur de larges ensembles de données relationnelles (tabulaires) réelles afin d'apprendre à reconnaître des schémas spécifiques. Si la personne concernée n'est jamais

en elle-même un objet d'intérêt, l'intégrité référentielle des données doit néanmoins être préservée. L'IA doit apprendre que la même entité impliquée dans la « situation A » est également impliquée dans l' « activité B », à travers différentes tables de base de données et événements. Des conditions d'utilisation spécifiques autorisent l'usage des données clients à cette fin, à condition qu'elles soient sous une forme anonymisée.

5.1 Les limites des technologies standard renforçant la protection de la vie privée (PET – *Privacy Enhancing Technologies*)

Les données originelles du service *cloud* contiennent des identifiants directs (noms, adresses électroniques) et des identifiants issus des bases de données internes. L'extraction de ces données aux fins d'entraînement de l'IA exige des techniques robustes de protection de la vie privée. Toutefois, les technologies standard renforçant la protection de la vie privée (PET) ne suffisent pas pour la détection d'anomalies.

L'application de la confidentialité différentielle avec un budget de confidentialité strict (epsilon),[25] ajoute un niveau important de bruit statistique, ce qui réduit une grande partie de l'utilité des données. La technique dite de k-anonymat[26] échoue pour la raison inverse. Il dissimule chaque personne dans une foule de personnes similaires, supprimant ainsi les valeurs atypiques mêmes dont un détecteur d'anomalies a besoin. Le prestataire a donc besoin d'une architecture différente.

5.2 Le hachage salé et la séparation stricte des environnements

Dès lors que l'identification directe est sans pertinence pour l'objectif d'entraînement, l'organisation remplace les identifiants internes par des pseudonymes. Afin de préserver la corrélation nécessaire entre les tables (intégrité référentielle) sans exposer les personnes concernées, le prestataire applique un hachage salé aux identifiants internes.

Surtout, la valeur de sel est stockée exclusivement au sein de l'environnement de production *cloud* hautement sécurisé.[27] Les données hachées sont ensuite transférées vers un environnement d'entraînement de l'IA strictement cloisonné. Les *data scientists* et les développeurs internes travaillant dans cet environnement d'entraînement n'ont absolument aucun accès à l'environnement *cloud* de production, aux données originelles ou à la valeur de sel.

5.3 Les schémas atteignent l'entraînement, mais le responsable du traitement conserve la clé

Ce cas d'usage présente une configuration factuelle plus complexe que les deux cas précédents. L'architecture préserve intentionnellement la corrélation au sein du jeu de données, et quiconque détient le sel peut techniquement inverser le hachage. Ces deux caractéristiques sont nécessaires au fonctionnement de la détection d'anomalies.

---

[25] ibid
[26] ibid
[27] Un sel secret unique se comporte comme une clé maîtresse. Il protège l'ensemble du jeu de données et concentre également l'ensemble du risque en un seul point. Une seule fuite de ce sel, par erreur, par un acteur interne, ou par la compromission du système de production, inverse instantanément tous les identifiants hachés.

Étape 1. La Cour indique (nous paraphrasons) que le CRU détient les informations supplémentaires permettant d'attribuer les commentaires à une personne, de sorte que les commentaires demeurent personnels pour le CRU malgré la pseudonymisation.[28] La Cour se penche ensuite sur Deloitte, société distincte à laquelle le CRU a transmis des commentaires pseudonymisés. Pour Deloitte, les commentaires peuvent être non personnels, mais uniquement si Deloitte ne peut lever les mesures pendant un traitement effectué sous son propre contrôle et uniquement si les mesures empêchent effectivement Deloitte d'attribuer les commentaires à une personne.[29]

Ce cas d'usage ne comporte aucune entité « Deloitte ». Les *data scientists* travaillent au sein de la même organisation, le prestataire conserve le sel dans son environnement de production et les données circulent entre deux entités d'une même société, non entre deux sociétés distinctes. Le point 77 de l'arrêt CRU ne trouve donc pas à s'appliquer. Le prestataire contrôle le traitement dans l'environnement d'entraînement et peut lever les mesures à tout moment, puisqu'il détient le sel. L'arrêt suppose un destinataire distinct traitant sous son propre contrôle, or il n'en existe aucun ic Le point 76 de l'arrêt CRU s'applique donc pleinement : le prestataire est en même temps le responsable du traitement qui pseudonymise les données et qui conserve la clé, de sorte que les données hachées et salées demeurent personnelles entre ses mains, environnement d'entraînement compris.

Le point 78 de l'arrêt CRU referme la dernière ouverture lorsqu'on le lit conjointement avec le test sur la disponibilité des moyens de réidentification. Le point 16 de l'arrêt SRB qualifie de donnée personnelle (ou donnée concernant une personne identifiable), une donnée pseudonymisée lorsque des informations supplémentaires permettent de l'attribuer à cette personne. La Cour relie ce considérant aux moyens raisonnablement susceptibles d'être utilisés pour lesquels le coût et le temps de l'identification sont des facteurs objectifs à prendre en compte. Pour ce prestataire, le test n'appelle aucune pondération particulière. Il conserve le sel actif dans son propre environnement de production et peut relier le hachage à l'identifiant sur demande. L'étape 1 oriente donc, à notre sens, vers une donnée personnelle.

Étape 2. Les données échouent dès l'étape 1, et le cadre s'arrête là. Par souci d'exhaustivité, nous analysons néanmoins deux critères, car ils montrent là où l'architecture réussit et là où elle fait défaut.

Cas d'usage 3 – Cadre d'évaluation CRU	
Critère	**Evaluation**
Séparation	La séparation entre l'environnement d'entraînement et la production est structurelle et cryptographique et elle est réelle. Trois couches tiennent le sel à l'écart des data scientists : des contrôles techniques, une frontière organisationnelle et un contrat. C'est la partie la plus solide de l'architecture et elle accomplit un travail de sécurité réel. Elle ne fait toutefois pas disparaître le sel pour le responsable du traitement, qui détient les deux environnements.

[28] Affaire C-413/23 P, point 76.
[29] Affaire C-413/23 P, point 77.

Irréversibilité	Ce critère, qui détermine le test de recoupement de l'avis de 2014 sur l'anonymisation, échoue. La présente étude considère que, sans accès au sel, la reconstitution de la valeur d'origine par force brute n'est pas raisonnablement réalisable au regard des ressources de calcul nécessaires. Nous estimons que cela vaut globalement pour les data scientists pris isolément. L'avis de 2014 n'apprécie cependant pas l'irréversibilité au regard d'une seule équipe au sein du responsable du traitement, et le prestataire conserve le sel actif en production pour ses propres finalités, de sorte qu'il peut relier le hachage à l'identifiant sur demande. La clé de réversion constitue une caractéristique permanente du système, non un risque résiduel Des données qu'un responsable du traitement peut relier à nouveau ne sauraient être considérées comme irréversiblement anonymisées.

Les données ne franchissent pas le seuil requis à la première étape de l'analyse, et la considération tenant à l'irréversibilité confirme cette conclusion. Nous estimons que les données d'entraînement hachées et salées constituent des données personnelles entre les mains du prestataire, donc très vraisemblablement non anonymes, au sens du considérant 26 du RGPD. Il ne s'agit pas d'un cas limite tranché par précaution. C'est une lecture directe des points 76 et 16 de l'arrêt CRU : le prestataire conserve la clé, donc le prestataire détient des données personnelles et l'environnement d'entraînement cloisonné n'est qu'une composante de ce même prestataire. Si le prestataire avait partagé les données avec un destinataire indépendant ne pouvant obtenir le sel, le point 77 de l'arrêt CRU trouverait à s'appliquer et les données pourraient très vraisemblablement être qualifiées d'anonymes entre les mains de ce destinataire.

Ce cas d'usage trouve toute sa place dans le débat plus large sur la protection des données personnelles. Il réunit la combinaison la plus solide qui soit, (1) une séparation cryptographique, (2) une scission organisationnelle et (3) un contrat contraignant pour les *data scientists*. Pourtant, il se range du côté des données personnelles, parce que le prestataire conserve le sel. C'est ici que nous rencontrons la limite que l'arrêt CRU trace autour d'un responsable du traitement qui détient la clé. Au-delà de cet arrêt, la jurisprudence offre peu d'orientation pour guider une telle architecture, qui se trouve donc en territoire inexploré. Seules des lignes directrices plus précises ou une norme de certification reconnue montreront aux développeurs jusqu'où une telle architecture peut véritablement se rapprocher de l'anonymat.

6. Cas d'usage 4 : Transmission de statistiques d'audience agrégées à une plateforme éditoriale

Cadre d'évaluation

Au regard du cadre d'évaluation présenté dans la Section 2, les rapports d'audience agrégés franchissent les deux étapes. Nous estimons qu'ils sont très vraisemblablement anonymes entre les mains de la plateforme éditoriale en tant que destinataire, tandis que le prestataire de mesure continue de détenir des données personnelles.

Résumé du cas. Ce cas a été fourni par une plateforme diffusant des contenus vidéo et d'actualité, à l'appui d'une mesure d'audience respectueuse de la protection données personnelles. Les éditeurs en ligne ont besoin de statistiques fiables sur leur audience, mais ces rapports sont produits à partir de données sous-jacentes qui peuvent initialement se rapporter à des utilisateurs individuels. Le cas d'usage montre comment l'agrégation, des flux de données à sens unique, des restrictions contractuelles et une séparation commerciale garantissent que l'éditeur ne reçoit que des statistiques au niveau de la population, sans accès aux données brutes, aux listes de panélistes, ni aux clés d'identification. Il en résulte que l'éditeur ne dispose d'aucun moyen réaliste d'individualiser, corréler ou inférer des informations sur une personne précise à partir des rapports. Une limitation subsiste. Une cellule démographique de petite taille, dépourvue d'un seuil minimal, laisse subsister un risque d'inférence d'appartenance pour un acteur disposant déjà d'informations relatives à une personne déterminée. Traiter de tels rapports comme non personnels, lorsque ces garanties sont effectives, peut favoriser en Europe une mesure d'audience fiable, qui sous-tend la monétisation des espaces publicitaires et l'amélioration des services.

Ce cas d'usage examine si les rapports de mesure d'audience agrégés, transmis par un prestataire de mesure spécialisé à un éditeur en ligne, constituent des données à caractère personnel entre les mains de l'éditeur, alors même que le prestataire de mesure a traité des données personnelles de visiteurs de sites et d'utilisateurs d'applications individuels pour produire ces rapports.

6.1 La collecte et le traitement des données par le prestataire spécialisé dans la mesure d'audience

Un prestataire spécialisé dans la mesure d'audience (le Prestataire) propose un service mensuel à un éditeur de médias en ligne (la Plateforme). Le Prestataire collecte les données, les traite, et transmet des rapports statistiques à la Plateforme. Ces rapports indiquent à la Plateforme combien de personnes ont visité ses pages, qui sont ces personnes en termes démographiques agrégés et comment l'audience de la Plateforme se compare à la population internet totale d'un État membre de l'UE.

Le flux de données comporte deux étapes et deux flux de données. À l'étape 1, le Prestataire collecte des données au niveau individuel via deux flux de données distincts. Des personnes se portant volontaires comme panélistes installent un logiciel de mesure sur leurs appareils et consentent au suivi de leur comportement de navigation sur ordinateurs, téléphones mobiles et tablettes (flux de données 1). En parallèle, et seulement *après* que l'utilisateur, c'est-à-dire le visiteur du site ou l'utilisateur de l'application, a donné son consentement, le Prestataire déploie un logiciel spécialisé sur les sites web et applications mobiles de la Plateforme (le SDK de mesure d'audience, ou *Software Development Kit*, une bibliothèque de code gérant des fonctions spécifiques). Ce logiciel suit des actions précises de l'utilisateur, telles que la lecture, la mise en pause ou l'arrêt d'une vidéo, et transmet ces données directement depuis le navigateur ou l'application de l'utilisateur vers les serveurs du Prestataire. Chaque transmission comprend également des informations sur le contenu visionné, telles que le titre et la durée de la vidéo, des informations de base sur l'appareil de l'utilisateur et un

enregistrement de ce à quoi il a consenti (flux de données 2).[30] Les données ne transitent jamais par la propre infrastructure de la Plateforme. Elles vont directement au Prestataire.

À l'étape 2, le Prestataire (a) combine les deux flux de données (flux 1 et 2), (b) corrige les biais de panel[31] et (c) extrapole les résultats afin de représenter la population internet totale d'un État membre de l'UE. Le résultat est un ensemble de rapports statistiques mensuels. Ces rapports agrégés constituent la seule chose que le Prestataire transmet à la Plateforme. La Plateforme n'a accès ni aux données sous-jacentes, ni au registre des panélistes, ni aux algorithmes de traitement, ni à aucun jeu de données intermédiaire.

6.2 L'agrégation des données comme principale garantie de protection de la vie privée

La principale technique de protection des données personnelles mise en œuvre dans ce cas d'usage est l'agrégation. L'agrégation regroupe des enregistrements individuels en décomptes et synthèses statistiques de groupe. Une fois exprimée au niveau de la population, la contribution individuelle disparaît dans le groupe. Il ne subsiste aucun enregistrement permettant d'individualiser, de corréler ou d'inverser le processus.

En pratique, le Prestataire croise les signaux d'un grand nombre d'utilisateurs en chiffres tels qu'un taux de couverture ou un nombre de visiteurs uniques pour un mois donné. Aucun comportement individuel ne produit, à lui seul, un chiffre. Chaque chiffre reflète l'activité d'une population substantielle. Les rapports ne comportent aucun enregistrement individuel, aucun quasi-identifiant et aucun point de donnée à partir duquel une personne physique pourrait être identifiée. Les rapports, à eux seuls, ne permettent à personne de reconstituer un individu à partir de ces chiffres.

L'agrégation, cependant, n'élimine pas à elle seule toute probabilité de réidentification. Une attaque par inférence d'appartenance[32] pose une question plus circonscrite. Plutôt que d'identifier qui est une personne, elle interroge si les données d'une personne précise ont contribué à un agrégat donné. Cela demeure une menace réaliste, même là où une réidentification complète ne l'est pas. En théorie, un adversaire disposant d'informations supplémentaires sur une personne précise pourrait tenter de détecter sa présence dans un rapport mensuel en comparant les résultats dans le temps ou entre cellules démographiques.

Dans ce cas d'usage, ce risque demeure faible en raison de l'architecture spécifique mise en œuvre.[33] Quelles que soient les données de première partie que la Plateforme détient du fait

---

[30] Les données collectées par le Prestataire dépendent du produit choisi par la Plateforme. Le produit de base, centré sur le site, mobilise les deux flux de données. Le flux de données 1 (panélistes) et le flux de données 2 (le SDK) fournissent ensemble l'adresse IP, la page visitée, le volume de visites et le type d'appareil, permettant à la Plateforme de recevoir des décomptes de visites mensuels et quotidiens moyens par canal, des ventilations sociodémographiques, et des classements d'audience. L'option centrée sur la vidéo va plus loin. Un traceur distinct, complémentaire au flux de données 2, collecte le titre de la vidéo, la durée de visionnage, le comportement de visionnage, l'adresse IP, et le volume de vidéos regardées. La Plateforme reçoit alors des décomptes de spectateurs uniques, le temps passé par vidéo, une répartition de l'audience par type d'appareil, des ventilations démographiques, et des classements.

[31] Voir le glossaire.

[32] ibid

[33] Cette lecture favorable vaut pour la Plateforme, qui ne peut interroger les rapports. Elle ne vaut pas de manière générale. En l'absence de confidentialité différentielle ou de taille minimale de cellule, une cellule démographique restreinte, telle qu'une tranche d'âge étroite dans une petite région, laisse subsister une possibilité d'inférence d'appartenance pour un acteur détenant des données supplémentaires sur une personne précise. À notre avis, une taille minimale de cellule supprimerait ce risque.

de l'exploitation de ses propres sites et applications, les rapports mensuels ne lui donnent rien à quoi comparer ces enregistrements. Chaque rapport comporte des décomptes au niveau de la population et des répartitions démographiques, sans enregistrement individuel ni identifiant. Le SDK transmet les signaux de mesure directement au Prestataire, de sorte que la Plateforme n'apprend jamais lesquels de ses visiteurs le Prestataire a comptabilisé. Les rapports sont également des livraisons mensuelles statiques, sans mécanisme d'interrogation. La technique standard d'attaque par inférence d'appartenance interroge un système de manière répétée, avec et sans une personne précise, pour détecter une différence. La Plateforme n'a aucun moyen de procéder ainsi. Sur des millions d'utilisateurs internet, la contribution de toute personne prise isolément à un chiffre mensuel est, à notre avis, pratiquement indétectable. La seule exception concerne une cellule démographique restreinte. Le paragraphe suivant traite de ce risque.

Une protection mathématique plus robuste consisterait à appliquer un bruit de confidentialité différentielle aux résultats ou à imposer des seuils minimaux de taille de cellule,[34] pour les ventilations démographiques. La confidentialité différentielle[35] ajoute un bruit statistique calibré de sorte que la présence ou l'absence d'une personne ne puisse être détectée à partir du résultat. Cela va plus loin que l'agrégation, car cela fournit une garantie mathématique plutôt qu'un obstacle pratique. Les seuils minimaux de taille de cellule suppriment toute cellule démographique comportant moins d'un nombre défini de personnes, empêchant l'inférence d'appartenance par combinaison de petits groupes. Lorsque le Prestataire applique l'une ou l'autre mesure, la protection devient mathématique plutôt qu'architecturale. Le cas d'usage ne dépend pas de ces mesures pour aboutir à sa conclusion, mais leur application réduirait davantage le risque résiduel et renforcerait l'évaluation globale de l'anonymisation.

La pseudonymisation est également présente, mais à un stade plus précoce. La couche de collecte centrée sur le site relie les signaux à des identifiants d'appareil et à des identifiants fondés sur des cookies plutôt qu'à des noms ou à des données directement identifiantes. La Plateforme ne reçoit jamais même cette couche pseudonymisée. L'agrégation l'absorbe entièrement avant que les rapports ne soient produits.

Quatre garde-fous convergents renforcent l'agrégation. Le premier est *structurel*. La Plateforme ne reçoit que les rapports agrégés. Les données au niveau individuel à partir desquelles ils ont été produits restent chez le Prestataire. Le SDK envoie des requêtes HTTP directement depuis le navigateur de l'utilisateur vers les serveurs du Prestataire, de sorte que les signaux bruts ne transitent jamais par l'infrastructure de la Plateforme. La séparation est intégrée dans l'architecture des données, et non simplement promise sur le papier.

Le deuxième garde-fou est *technique*. Réidentifier une personne quelconque à partir des rapports exigerait que la Plateforme obtienne quatre éléments qu'elle ne possède pas. Il lui faudrait (1) les signaux bruts d'appareil et comportementaux transmis par le SDK, (2) le registre des panélistes, (3) les algorithmes d'hybridation et d'extrapolation, et (4) la correspondance entre signaux d'appareil et panélistes identifiés. Aucun de ces éléments n'atteint la Plateforme. Le SDK envoie des requêtes HTTP directement depuis le navigateur de l'utilisateur vers les serveurs du Prestataire. Les données ne transitent jamais par la propre infrastructure de la Plateforme.

[34] Voir le glossaire.
[35] ibid

Le troisième garde-fou est *contractuel*. Il est interdit à la Plateforme de procéder à une ingénierie inverse des rapports ou de réidentifier des personnes à partir de ceux-ci. L'accord contractuel l'empêche de modifier, fusionner, transformer ou combiner les rapports avec toute autre donnée sans le consentement écrit préalable du Prestataire. Ces interdictions emportent un risque juridique significatif et dissuadent toute tentative en ce sens.

Le quatrième garde-fou est *commercial*. La méthodologie du Prestataire, la composition de son panel et ses algorithmes d'extrapolation constituent ses principaux secrets d'affaires. Les divulguer à un éditeur-client porterait atteinte à la neutralité, vis-à-vis des éditeurs, d'une mesure certifiée dont dépend l'ensemble du marché, exposerait les données des panélistes et créerait un risque concurrentiel sérieux. Même si une Plateforme cherchait à obtenir ces informations, le Prestataire disposerait de toutes les incitations structurelles pour refuser. Les « moyens raisonnablement susceptibles d'être utilisés » au sens du considérant 26 du RGPD doivent tenir compte non seulement de ce qu'un destinataire pourrait techniquement tenter, mais également de ce qu'un tiers détenant les informations nécessaires accepterait réalistement de fournir. En l'espèce, une telle coopération est commercialement inconcevable. Les classements ajoutent une couche supplémentaire. De nombreux classements d'audience circulent publiquement sur le marché ; une Plateforme classée cinquième sait déjà quelle autre Plateforme se situe au-dessus et en dessous d'elle. La valeur marginale d'une tentative de réidentification de personnes individuelles à partir de rapports en partie publics est, à notre avis, négligeable. Une telle tentative présente également un caractère contre-productif. La Plateforme a commandé ces rapports précisément parce que le marché leur fait confiance. Tout soupçon de manipulation détruirait cette confiance et anéantirait l'investissement consenti. Les rapports couvrent en outre une période brève et définie, typiquement un seul mois. Le temps qu'une tentative de réidentification puisse être conçue et exécutée, les données auraient perdu toute pertinence commerciale.

Un point supplémentaire mérite attention. Le Prestataire supprime ou anonymise ses données brutes conformément à son propre calendrier de conservation de données, mais la Plateforme n'a aucune visibilité sur ce calendrier. Selon une lecture extensive de la notion de donnée personnelle, les rapports pourraient être qualifiés de données personnelles le jour de leur livraison, puis de données non personnelles des mois plus tard, une fois que le Prestataire a supprimé les enregistrements sous-jacents, sans que la Plateforme ne soit jamais informée de ce changement, dans un sens comme dans l'autre. Il s'agirait là d'une obligation de conformité impossible à satisfaire. La Plateforme serait tenue d'appliquer les règles du RGPD à des données dont elle ne peut déterminer, ni influencer le statut juridique. L'arrêt CRU résout cette absurdité en ancrant le caractère identifiable aux moyens dont dispose le destinataire, plutôt qu'aux capacités théoriques d'un tiers détenant des données séparées et inaccessibles.

6.3 Quatre garde-fous convergent pour placer les rapports hors du champ d'application du RGPD pour la Plateforme

L'évaluation se place du point de vue de la Plateforme, car le Prestataire détient le registre des panélistes, les signaux bruts et les correspondances d'identifiants, de sorte que les données demeurent personnelles entre ses mains.

Étape 1. L'arrêt CRU pose deux conditions cumulatives pour le destinataire, ici la Plateforme. La première interroge la capacité de la Plateforme à lever les mesures pendant son traitement. Nous estimons qu'elle ne le peut pas. L'agrégation, l'hybridation des panels et l'extrapolation sont effectués par le Prestataire ; la Plateforme ne reçoit que les rapports mensuels finalisés.

Elle ne dispose ni de signaux bruts, ni de registre de panélistes, ni d'algorithmes, ni d'aucun moyen d'interroger les données sous-jacentes et ne peut donc défaire une agrégation qu'elle n'a jamais réalisée.

La seconde condition interroge si les mesures empêchent effectivement la Plateforme d'attribuer un chiffre à une personne physique, y compris par recoupement avec d'autres facteurs. Nous estimons qu'elles le font. Chaque chiffre regroupe l'activité d'une population substantielle, le SDK transmet les signaux directement depuis le navigateur de l'utilisateur vers le Prestataire, et la Plateforme n'apprend jamais lesquels de ses visiteurs le Prestataire a comptabilisé. Un décompte mensuel ne contient aucun enregistrement individuel à recouper, quelles que soient les données propres détenues par la Plateforme.

Un scénario appelle une attention particulière. Une cellule démographique de faible effectif, telle qu'une tranche d'âge étroite dans une petite région, peut se distinguer, et une agrégation unique peut rendre à nouveau possible une inférence d'appartenance pour quiconque détient des données sur une personne précise. Nous ne fondons pas le verdict sur l'absence de telles données chez la Plateforme. Nous le fondons sur la condition de taille minimale de cellule de l'étape 2. Lorsque le Prestataire supprime les cellules restreintes, la seconde condition est satisfaite, et à notre avis le test CRU oriente vers l'anonymat pour ce destinataire.

Étape 2, le seuil de précaution de l'avis de 2014 du CEPD sur l'anonymisation. Nous soumettons ensuite les rapports aux critères plus stricts de l'avis de 2014 sur l'anonymisation.

Cas d'usage 4 – Cadre d'évaluation CRU	
Critère	**Evaluation**
Séparation	La Plateforme ne reçoit que les rapports agrégés. Les enregistrements de panels, les signaux bruts d'appareil et les correspondances d'identifiants restent chez le Prestataire, alors que le SDK envoie ses requêtes directement depuis le navigateur de l'utilisateur vers les serveurs du Prestataire, de sorte que les signaux bruts ne touchent jamais l'infrastructure de la Plateforme. Le flux de données crée la séparation, non une promesse sur le papier.
Irréversibilité	La Plateforme ne peut inverser l'agrégation. Elle ne dispose ni des données brutes en entrée, ni du registre des panélistes, ni des algorithmes d'hybridation qui transforment les données de panel et les mesures de site en estimations de population. Une fois qu'un chiffre regroupe des millions de contributions, aucun enregistrement ne subsiste que la Plateforme pourrait défaire.
Réidentification	La probabilité de réidentifier une personne physique à partir des rapports demeure très vraisemblablement insignifiante. Chaque chiffre regroupe l'activité d'une population substantielle, de sorte qu'aucun enregistrement individuel ne subsiste dans le résultat. Réidentifier une seule personne exigerait les signaux bruts, le registre des panélistes, les algorithmes d'extrapolation et la correspondance entre signaux d'appareil et panélistes. La Plateforme ne reçoit aucun de ces éléments.

Individualisation	Les rapports ne comportent aucun enregistrement au niveau individuel, ni aucun quasi-identifiant. Il n'est pas plausible d'individualiser une personne physique à partir d'un décompte au niveau de la population.
Corrélation	La Plateforme ne détient ni registre de panélistes, ni données brutes de tag, ni correspondance d'identifiants. Un chiffre au niveau de la population ne contient aucun enregistrement permettant de le corréler avec un autre jeu de données.
Inférence	Les rapports décrivent des distributions sur de larges segments et ne disent rien du comportement d'une personne précise. Le risque résiduel se situe ici plutôt qu'ailleurs. Lorsqu'une ventilation démographique produit une cellule restreinte et que le Prestataire n'applique ni confidentialité différentielle, ni seuil minimal de taille de cellule, une inférence d'appartenance demeure possible pour un acteur détenant des données supplémentaires sur une personne précise. Un seuil minimal de taille de cellule répond directement à cette hypothèse. Lorsque le Prestataire supprime les cellules en deçà d'un minimum fixé, ce risque disparaît et le verdict de très vraisemblablement anonyme vaut pour les ventilations. En l'absence de cette suppression, les ventilations démographiques restreintes peuvent demeurer des données personnelles.

Les rapports franchissent les deux étapes pour le destinataire. Nous estimons que les rapports de mesure d'audience agrégés sont très vraisemblablement anonymes entre les mains de la Plateforme, au sens du considérant 26 du RGPD et de l'arrêt CRU, et qu'ils demeurent personnels entre les mains du Prestataire.

Deux limites encadrent ce verdict. Premièrement, il dépend d'une agrégation réelle portant sur de vastes groupes de personnes. Lorsqu'une cellule démographique est restreinte et que le Prestataire ne fixe aucune taille minimale de cellule et qu'il n'ajoute par ailleurs aucun bruit de confidentialité différentielle, une inférence d'appartenance demeure possible. Le verdict relatif aux ventilations démographiques ne vaut donc que là où le Prestataire supprime les cellules en deçà d'un minimum fixé et peut le démontrer sur demande. Deuxièmement, le verdict est propre au destinataire. Il vaut pour la Plateforme, dont les données propres ne peuvent se connecter aux chiffres au niveau de la population. Il ne s'applique pas à un destinataire disposant de moyens raisonnablement susceptibles d'identifier un contributeur. Chaque nouveau destinataire appelle sa propre évaluation.

7. Cas d'usage 5 : Entraînement de modèles de prédiction publicitaire sans données des utilisateurs individuels

Cadre d'évaluation

Au regard du cadre exposé en Section 2, le jeu de données d'entraînement synthétiques franchit les deux étapes. Nous estimons qu'il est très vraisemblablement anonyme entre les mains du prestataire publicitaire, tandis que les enregistrements source demeurent des données personnelles.

Résumé du cas. Ce cas a été fourni par un prestataire européen de publicité en ligne, à l'appui d'une publicité respectueuse des données personnelles. Pour entraîner des modèles d'IA prédisant si des utilisateurs sont susceptibles de cliquer sur des publicités, l'organisation ne s'appuie pas sur des enregistrements de clics individuels. Le cas d'usage montre comment la pseudonymisation, l'agrégation, le bruit statistique et les données synthétiques peuvent être croisés afin que l'entraînement du modèle utilise des schémas observés au niveau du groupe plutôt que des données identifiables au niveau de l'utilisateur. Il en résulte que le modèle peut apprendre des signaux publicitaires utiles sans exposer qui a cliqué, sans relier des enregistrements à des individus, ni inférer le comportement d'une personne précise. Une limite demeure toutefois. Le verdict repose sur un budget de confidentialité restreint et suivi, ainsi que sur des groupes suffisamment importants pour empêcher l'identification de toute personne individuelle. Traiter les données d'entraînement synthétiques comme non personnelles, lorsque ces garanties sont effectives, peut favoriser en Europe le développement d'une IA respectueuse de la vie privée tout en permettant l'innovation dans la publicité et dans d'autres secteurs innovant par la donnée.

Ce cas d'usage retrace comment un prestataire de publicité en ligne a utilisé quatre technologies renforçant la protection de la vie privée (PETs) afin de réduire la probabilité de réidentification à chaque étape de l'entraînement du modèle d'IA. La pseudonymisation, l'agrégation, la confidentialité différentielle et les données synthétiques concernent chacune une vulnérabilité distincte. Ensemble, elles abaissent le risque résiduel à un niveau compatible avec le cadre d'évaluation.

Un prestataire de publicité en ligne (l'Organisation) entraîne des modèles d'IA pour formuler des prédictions concernant des utilisateurs individuels, dans une logique de respect de la vie privée dès la conception. Ce cas vise à prédire la probabilité qu'un utilisateur clique sur une publicité en ligne. Pour entraîner un tel modèle, l'Organisation pourrait utiliser des données reliant des caractéristiques individuelles à des résultats individuels. Par exemple, une personne cliquant sur des chaussures de sport sur un site marchand pourrait apprécier les vêtements de sport. Toutefois, l'Organisation n'a pas accès à des données de résultat individuelles. Elle ne peut savoir si un utilisateur précis a cliqué. Elle n'a accès qu'à des totaux agrégés, tels que le nombre de clics au sein d'un (vaste) groupe d'utilisateurs.

Ce cas d'usage n'est pas pertinent uniquement pour la publicité en ligne. Le même défi se retrouve dans d'autres secteurs. Par exemple, un chercheur en médecine pourrait vouloir prédire si un patient est atteint d'une maladie précise, mais ne peut fonder cette prédiction que sur le nombre total de cas dans une population (c'est-à-dire l'ensemble de la population française). Dans les deux cas, les données au niveau individuel ne sont pas accessibles pour des raisons de protection des données personnelles.

7.1 Le remplacement des identifiants directs par des identifiants hachés ou aléatoires

L'Organisation part d'un jeu de données d'enregistrements pseudonymisés. Ces enregistrements contiennent des données non personnelles, telles que le type de campagne publicitaire ou l'emplacement de l'annonce. Ils incluent également des enregistrements hachés. Le hachage est un procédé mathématique consistant à brouiller une information, telle qu'une adresse électronique ou un numéro de téléphone, en une chaîne de longueur fixe de

lettres et de chiffres difficilement réversible. Un exemple d'équivalent haché d'un identifiant « 13278a5c-3997-4b97-826d-19609eecb975 » contenu dans un cookie est « c68e128c46549947416241878e16e9db811da0c99e7431d6982c56481008dc6c ». Ce type de données pseudonymisées comporte encore trois voies distinctes de réidentification.

Premièrement, un acteur malveillant peut individualiser une personne en recoupant plusieurs quasi-identifiants, même hachés. Un quasi-identifiant est une information qui ne permet pas d'identifier une personne à elle seule, mais qui permet de le faire si elle est croisée avec d'autres informations. Un code postal seul ne vous identifie pas. Votre date de naissance seule ne vous identifie pas. Mais le croisement d'un code postal avec une date de naissance partielle peut réduire un jeu de données à une seule personne. Deuxièmement, un adversaire peut corréler des enregistrements pseudonymisés avec des jeux de données externes contenant des informations directement identifiantes et ainsi recouvrer l'identité d'une personne. Troisièmement, un adversaire peut inférer les caractéristiques ou comportements d'une personne précise à partir de schémas présents dans les données, même sans identification directe.

Les données pseudonymisées, à elles seules, n'éliminent pas ces risques. L'Organisation applique donc deux technologies renforçant la protection de la vie privée supplémentaires aux données d'entraînement, et en utilise une troisième pendant l'entraînement du modèle.

7.2 Le regroupement des enregistrements individuels en comptages agrégés supprime les données individuelles

L'Organisation regroupe des enregistrements pseudonymisés individuels, provenant de nombreux utilisateurs, en comptage au niveau du groupe. Ces décomptes enregistrent combien d'utilisateurs partagent une caractéristique donnée et combien ont obtenu le résultat recherché, tel qu'un clic sur une publicité. L'enregistrement pseudonymisé disparaît. Un acteur malveillant souhaitant individualiser ou corréler une personne précise doit inverser l'agrégation. Cela exige une attaque par force brute sur l'ensemble des combinaisons possibles. À grande échelle, cela est computationnellement déraisonnable. L'attaquant doit également connaître la logique d'agrégation, que l'Organisation ne divulgue pas.

7.3 L'ajout de bruit statistique protège la contribution de chaque personne

Considérez le bruit statistique comme une imprécision délibérée. Plutôt que d'indiquer qu'exactement 47 personnes dans un code postal ont entre 30 et 40 ans, le système pourrait indiquer 43 ou 51. Le décompte réel varie d'un petit montant aléatoire à chaque fois. Cet aléa change entièrement le calcul pour un attaquant. L'Organisation ajoute un bruit statistique, au-delà d'un seuil défini, aux décomptes agrégés. Le niveau de bruit est régi par un paramètre de confidentialité configurable (epsilon),[36] qui détermine dans quelle mesure les données d'une personne peuvent influencer un résultat donné. L'Organisation fixe ce paramètre à un niveau offrant un déni plausible significatif pour chaque contribution individuelle. Dans l'exemple de la tranche d'âge démographique, le décompte rapporté pour un groupe donné peut varier de plus de deux années d'amplitude d'âge. La contribution de toute personne à un agrégat donné devient indétectable. Nul ne peut raisonnablement confirmer si les données

[36] ibid

d'une personne précise ont contribué à un agrégat donné. Cela bloque les attaques par inférence visant la présence ou les caractéristiques d'une personne dans le jeu de données.

7.4 La reconstruction des signaux d'entraînement à partir des décomptes agrégés évite le recours à des données au niveau de l'utilisateur

L'Organisation génère des données synthétiques en reconstruisant des signaux d'entraînement au niveau de l'utilisateur à partir de décomptes de groupe, sans utiliser d'enregistrements pseudonymisés. L'idée est de générer des données fictives mais plausibles, se comportant comme des données réelles, sans exposer qui que ce soit. Lorsque le résultat recherché n'est pas disponible au niveau d'un utilisateur isolé, l'Organisation ne peut entraîner directement le modèle. Par exemple, elle sait qu'un utilisateur sur 1 000 a cliqué sur une publicité. Elle ne sait pas lequel. L'Organisation utilise un modèle probabiliste pour générer des enregistrements individuels synthétiques cohérents avec les décomptes de groupe observés. Cela imite la manière dont des données des utilisateurs individuels auraient pu être générées. L'Organisation entraîne ensuite le modèle d'IA sur ce jeu de données synthétiques. Elle n'utilise jamais d'enregistrements individuels.

7.5 Le jeu de données synthétiques au regard de l'évaluation en deux étapes

L'architecture à quatre couches résiste aux techniques classiques de réidentification, y compris la combinaison de quasi-identifiants et la corrélation avec des jeux de données externes. Les données synthétiques garantissent que l'entraînement du modèle n'exige jamais l'accès à des enregistrements individuels. Après application des quatre technologies de protection de la vie privée, le profil de risque résiduel diffère substantiellement de celui du jeu de données pseudonymisées d'origine.

Ce cas d'usage illustre une architecture en couches de technologies de protection de la vie privée. Chaque technologie annule une menace de réidentification précise. Évaluée au regard du cadre d'évaluation, cette architecture en couches répond directement aux trois critères centraux d'anonymisation. Elle tient également compte des moyens raisonnablement susceptibles d'être utilisés.

Étape 1. Le jeu de données synthétiques constitue un objet différent. Le cas d'usage 3 conservait un sel secret permettant d'inverser chaque hachage sur demande. Ce processus fonctionne dans un seul sens et ne conserve aucune clé de ce type, de sorte que l'Organisation ne peut relier un enregistrement synthétique à une personne. Le considérant 26 du RGPD traite comme anonymes les données que personne ne peut attribuer à une personne, responsable du traitement compris. Pour le jeu de données synthétiques, l'étape 1 oriente vers l'anonymat, tandis que les enregistrements source demeurent personnels.

Étape 2. Un responsable du traitement unique qui anonymise sa propre production est exactement le cas où la précaution se justifie pleinement ; nous soumettons donc le jeu de données synthétiques aux critères plus stricts avant d'accepter le résultat.

Cas d'usage 5 – Cadre d'évaluation CRU	
Critère	**Evaluation**
Séparation	Le jeu de données synthétiques ne contient aucun enregistrement pseudonymisé. L'agrégation supprime l'enregistrement individuel avant même le début de la génération et les enregistrements synthétiques ne portent aucun identifiant issu de la source. La séparation est intégrée au processus, et non ajoutée après coup.
Irréversibilité	Ce critère détermine le test de recoupement. Ici, il est satisfait. La transformation ne fonctionne que dans un seul sens. L'agrégation détruit l'enregistrement, la confidentialité différentielle ajoute un bruit que l'Organisation ne peut soustraire et les données synthétiques rompent le lien avec les utilisateurs réels, de sorte que l'Organisation ne peut relier un enregistrement synthétique à une personne. L'étape de confidentialité différentielle porte cette conclusion, et ne la porte que tant que le budget de confidentialité (epsilon) demeure restreint et suivi à travers chaque publication.
Réidentification	La probabilité de réidentifier une personne physique demeure très vraisemblablement insignifiante. Un attaquant devrait simultanément inverser l'agrégation, soustraire le bruit et déjouer la génération synthétique. Les enregistrements eux-mêmes ne correspondent à aucune personne réelle.
Individualisation	Les enregistrements synthétiques ne décrivent aucune personne réelle. Un attaquant qui isole un enregistrement synthétique isole un enregistrement inventé, de sorte que l'opération n'atteint aucune personne physique.
Corrélation	Le jeu de données ne contient aucun enregistrement réel à comparer avec une source externe. Un enregistrement synthétique ne se relie à rien en dehors du modèle.
Inférence	Le risque résiduel se situe ici, plutôt qu'ailleurs. Le bruit statistique confère à chaque personne un déni plausible quant à une contribution à tout décompte et l'inférence d'appartenance visant une personne réelle demeure bloquée tant que le budget tient. Une cellule démographique de petite taille mérite néanmoins surveillance, car la protection fournie est statistique, et non pas absolue.

Les quatre technologies renforçant la protection de la vie privée opèrent en séquence, et chacune referme une brèche laissée ouverte par la précédente. La pseudonymisation supprime les identifiants directs, l'agrégation supprime l'enregistrement individuel, le bruit protège la contribution de chaque personne et les données synthétiques rompent le lien entre les entrées d'entraînement et les utilisateurs réels. La combinaison de ces quatre couches superposées ne laisse subsister aucun enregistrement individuel à quelque étape que ce soit de l'entraînement du modèle. Le risque résiduel se situe en deçà du seuil que le cadre d'évaluation traite comme anonyme, en tenant compte, à notre avis, des moyens raisonnablement susceptibles d'être utilisés.

Une objection mérite une réponse directe. L'avis de 2014 du CEPD sur l'anonymisation énonce que la confidentialité différentielle ne modifie pas les données d'origine. Tant que le responsable du traitement conserve ces données, il peut identifier des personnes dans les résultats bruités et ces résultats demeurent des données personnelles.[37] L'Organisation conserve ses enregistrements source, ce qui, à première vue, irait dans le sens contraire. La mise en garde porte toutefois sur des vues bruitées d'enregistrements réels, où chaque résultat décrit encore des personnes réelles. Les enregistrements synthétiques ne décrivent personne. Le générateur les construit uniquement à partir de décomptes de groupe, jamais à partir d'un enregistrement pseudonymisé, de sorte que les enregistrements source n'ouvrent aucune voie de retour. Le budget de confidentialité suivi plafonne ensuite ce que les décomptes peuvent révéler sur toute personne prise isolément. Nous lisons cet avis comme une mise en garde contre les raccourcis, non comme un obstacle à un processus qui supprime l'enregistrement avant même le début de l'entraînement.

Nous estimons que le jeu de données d'entraînement synthétiques est très vraisemblablement anonyme entre les mains de l'Organisation, au sens du considérant 26 du RGPD, tandis que les enregistrements source demeurent personnels entre ces mêmes mains. Cette conclusion repose sur le caractère véritablement irréversible de la transformation, non sur une promesse de conserver une clé sous clé.

Le verdict vaut pour le jeu de données synthétiques sous réserve d'un budget de confidentialité restreint et de tailles minimales de cellule.[38] Il ne s'étend pas aux enregistrements source, qui demeurent personnels.

8. Conclusions. Ce que les cas d'usage révèlent sur les données pseudonymisées en pratique

Les cinq cas d'usage relèvent de secteurs très différents. L'un transmet des *prompts* à un modèle d'IA générative. Un autre transforme des données de capteurs routiers en carte de trafic en temps réel. Un troisième entraîne une détection d'anomalies sur des données de métier relationnelles, mais le prestataire conserve la clé. Les deux derniers suivent la même logique, transposée à leur propre contexte : un éditeur qui ne reçoit que des rapports d'audience mensuels et un publicitaire qui entraîne ses modèles sur des enregistrements synthétiques bruités.

[37] Avis de 2014 sur l'anonymisation, section 3.1.3. « Il convient toutefois de préciser que les techniques de confidentialité différentielle ne modifient pas les données d'origine et que, par conséquent, tant que ces données d'origine subsistent, le responsable du traitement est en mesure d'identifier des personnes dans les résultats des requêtes de confidentialité différentielle, compte tenu de l'ensemble des moyens raisonnablement susceptibles d'être utilisés. De tels résultats doivent également être considérés comme des données à caractère personnel. » Voir également la section 3.1.3.3 sur les limites de la confidentialité différentielle.

[38] Les données synthétiques héritent des propriétés statistiques des décomptes dont elles procèdent. Une cellule qui ne couvre qu'une seule personne, ou seulement quelques-unes, ne décrit pas un groupe. Elle décrit cette personne. Le générateur peut alors reproduire cette combinaison rare d'attributs dans un enregistrement synthétique. Ce risque résiduel relève de la divulgation d'attribut plutôt que de l'inférence d'appartenance, et le seul budget de confidentialité ne suffit pas à l'éliminer. Des tailles minimales de cellule et la suppression des petits décomptes doivent combler cette brèche avant même le début de la génération.

8.1 Le dénominateur commun des cinq cas d'usage

Dans chaque cas, l'anonymisation repose sur davantage qu'une seule technique. Le prestataire d'intelligence juridique ne se contente pas d'expurger les *prompts*. Il signe également un accord de non-conservation et s'appuie sur l'incapacité du fournisseur d'IA à conserver quoi que ce soit au-delà de l'inférence. Le constructeur automobile ne se contente pas de retirer le NIV haché. Il retarde également les heures d'arrivée, supprime sa propre copie et ne transmet que des observations par segment routier. L'éditeur de logiciel métier hébergé dans le *cloud* ne se contente pas de hacher les identifiants clients. Il verrouille également le sel dans un environnement distinct, hors de portée de l'équipe de développement. Le prestataire de mesure d'audience ne se contente pas d'agréger. Il retient également le registre des panélistes, les algorithmes d'hybridation, la correspondance entre appareils et panélistes et adosse le tout à un contrat. Le dispositif de prédiction publicitaire superpose pseudonymisation, agrégation, bruit et données synthétiques, chaque couche comblant une brèche laissée par l'étape précédente.

Les conclusions ne vont pas toutes dans le même sens selon les cas d'usage. Nous estimons que quatre des cinq cas atteignent l'anonymat, mais uniquement pour le destinataire et uniquement au regard des faits. Les *prompts* expurgés sont très vraisemblablement anonymes pour le fournisseur d'IA, les cartes de trafic pour le prestataire tiers dès lors qu'un seuil minimal de contributeurs est respecté, les rapports d'audience pour la plateforme éditoriale, ainsi que les données d'entraînement synthétiques pour le prestataire de services publicitaires qui les construit. Dans chacun de ces quatre cas, les données demeurent personnelles pour la partie qui détient les enregistrements source. Le cas ne remplissant pas tous les critères constitue l'exception. Ce seul contre-exemple revêt une importance égale à celle des quatre cas concluants, dans la mesure où il aide à préciser les critères précis permettant de déterminer à quel moment des données pseudonymisées deviennent des données anonymes.

L'approche relative consacrée par la CJUE tient à une question simple. Le destinataire dispose-t-il de moyens raisonnables d'identifier des personnes ? Non pas si quelqu'un, quelque part, disposant de ressources illimitées et d'intentions malveillantes, pourrait en théorie reconstituer une identité. Les cas d'usage montrent pourquoi cette distinction importe. Du point de vue de Deloitte en tant que destinataire, les données sont qualifiées d'anonymes. Du point de vue du CRU en tant que responsable du traitement, les données sont personnelles, car le CRU détient l'ensemble des informations nécessaires pour identifier les auteurs. Dans le même temps, les cas d'usage montrent que la probabilité de réidentification ne saurait être écartée entièrement.

8.2 Perspectives

Les cas d'usage fournissent un cadre concret pour une application uniforme des exigences juridiques et pour déterminer avec davantage de clarté à quel moment des données détenues par une organisation sont qualifiées de données personnelles ou de données anonymes au sens de l'article 4 du RGPD. Les critères destinés à standardiser l'évaluation devraient, à notre avis,

1) Récompenser la séparation architecturale ;
2) Reconnaître la convergence de plusieurs garde-fous ;

3) Évaluer le risque à l'aune des capacités réelles de l'organisation ;
4) Traiter les contraintes contractuelles et commerciales comme des facteurs renforçant les garanties plutôt que comme des garanties autonomes ;
5) Permettre une appréciation des risques tenant compte des faits et circonstances propres à chaque cas d'usage. Par exemple, les pièces jointes contenant des données à caractère personnel appellent toujours une vigilance particulière, même lorsque les *prompts* qui les accompagnent n'en appellent pas. De même, un trafic peu dense sur des routes rurales justifie toujours un seuil minimal de contributeurs afin d'empêcher l'individualisation ;
6) Garantir que l'évaluation soit contextuelle et propre à chaque destinataire, car un même jeu de données peut être une donnée personnelle entre certaines mains et anonyme entre d'autres ;
7) La protection par couches mérite une reconnaissance explicite, car la force de ces cas tient à la manière dont les couches interagissent entre elles, et non à une méthode unique.

Nous énonçons clairement les limites de ces sept critères. Ils reposent sur cinq cas d'usage. Chacun provient d'une seule entreprise européenne ayant un intérêt dans le résultat. Cinq as d'usage constituent un échantillon réduit. Ce nombre réduit favorise en outre le succès, puisque quatre des cinq cas d'usage atteignent l'anonymat au regard du cadre d'évaluation, et qu'aucune entreprise ne propose volontairement une architecture dont elle anticipe l'échec. Nous présentons donc ces critères comme des hypothèses de travail plutôt que comme des conclusions arrêtées. Ils sont suffisamment précis pour constituer un point de départ. Des tests d'attaque indépendants, un contrôle des autorités de contrôle ou un projet pilote de certification au titre de l'article 42 du RGPD pourraient chacun les confirmer ou les corriger. Nous accueillerions favorablement les trois.

En conclusion, le présent rapport ne constitue pas un avis juridique et nous ne sommes pas une autorité de contrôle. Nous sommes des praticiens de la protection des données personnelles appliquant un cadre d'évaluation dans un esprit de précaution. Lorsqu'un risque apparaît ou s'accroît dans des circonstances précises, la réponse appropriée consiste en une atténuation claire ou en un risque résiduel documenté, non en une présomption générale de conformité.

Nous encourageons la Commission européenne, les autorités de contrôle et les autres parties prenantes concernées à utiliser ces cas d'usage comme fondement pour définir des critères clairs et harmonisés, permettant aux organisations de déterminer, avec une plus grande sécurité juridique, quand des données relèvent ou non du champ d'application du RGPD. Parallèlement, ce travail devrait être complété, en coopération étroite avec l'industrie, par le développement d'instruments pratiques tels que des codes de conduite et des cadres de certification, susceptibles d'opérationnaliser ces critères, de favoriser une application cohérente entre les secteurs et de récompenser les pratiques d'anonymisation robustes. Une telle approche peut garantir un niveau élevé de protection pour les personnes tout en favorisant l'innovation et un usage responsable des données à grande échelle.

Nous remercions chaque lecteur qui a pris le temps d'aller jusqu'au bout de ce rapport et espérons que ces exemples seront utiles à ceux qui œuvrent à l'élaboration de critères harmonisés et de lignes directrices plus claires.

GLOSSAIRE DES TERMES TECHNIQUES

IA agentique
L'IA agentique désigne des systèmes capables de planifier et d'exécuter de manière autonome des tâches en plusieurs étapes, sans intervention humaine à chaque stade. Contrairement à un agent conversationnel qui répond à une question à la fois, un système agentique enchaîne des opérations, accède à plusieurs sources de données, recourt si nécessaire à des outils spécialisés et agit sur la base des résultats obtenus. Cela importe pour le risque de réidentification, car l'échelle et l'automatisation modifient ce qui constitue un moyen raisonnablement susceptible d'être utilisé au sens du considérant 26 du RGPD.

Confidentialité différentielle
La confidentialité différentielle ajoute un bruit aléatoire calibré aux résultats statistiques. Le résultat paraît exact au niveau de la population, mais rend impossible de déterminer si les données d'une personne précise ont contribué à un chiffre donné.

Epsilon
Epsilon est le paramètre qui contrôle la quantité de bruit que la confidentialité différentielle injecte dans un résultat. Une valeur plus faible signifie une protection plus forte et une distorsion plus importante des chiffres sous-jacents.

k-anonymat
Le k-anonymat dissimule chaque personne dans une foule de personnes similaires. Il regroupe les enregistrements selon leurs *quasi-identifiants*, les éléments qui, combinés, permettent de cibler une personne, tels que le code postal, l'âge ou la fonction occupée, de sorte que chaque combinaison apparaisse au moins k fois. Pour y parvenir, il grossit la granularité des données, transformant par exemple un âge exact en tranche d'âge.

Attaque par inférence d'appartenance
Une attaque par inférence d'appartenance pose une question plus étroite que la réidentification. Plutôt que de demander qui est une personne, elle demande si les données d'une personne précise ont contribué à un agrégat donné. Cette distinction importe, car cette attaque fonctionne même contre des résultats correctement anonymisés.

Seuil minimal de taille de cellule
Un seuil minimal de taille de cellule supprime tout résultat statistique reposant sur moins d'un nombre défini de personnes. En son absence, un rapport portant sur un groupe démographique de très petite taille peut effectivement désigner des personnes précises. Par exemple, un petit groupe portant des chemises rouges au sein d'une foule en chemises bleues. Les autorités de contrôle commencent à traiter la probabilité d'individualisation comme une exigence de base.

Biais de panel et extrapolation
Un panel de mesure est un groupe de personnes qui ont accepté le suivi de leur comportement. De tels panels ne reflétant jamais parfaitement la population plus large, un écart systématique apparaît entre l'échantillon et la population qu'il représente. L'extrapolation corrige cet écart par une modélisation statistique, permettant à un prestataire de produire des estimations au niveau de la population à partir d'un échantillon limité.

Quasi-identifiant
Un quasi-identifiant est une donnée qui, prise isolément, n'identifie aucune personne, mais qui peut permettre de le faire lorsqu'elle est combinée à d'autres informations. Un code postal seul ne révèle rien. Combiné à une date de naissance, il peut réduire un jeu de données à une seule personne. La probabilité d'identification ne réside pas dans un champ unique, mais dans ce qu'un acteur malveillant peut assembler à partir de plusieurs d'entre eux.

Agrégation collective
L'agrégation collective survient lorsque tant d'appareils déclarent simultanément des observations pour le même lieu qu'aucune contribution d'appareil pris isolément ne se distingue. Le volume de données qui se chevauchent rend l'attribution pratiquement impossible. Des schémas singuliers, tels qu'une route rurale calme la nuit, affaiblissent cet effet et justifient des garanties supplémentaires.

Données synthétiques
Les données synthétiques sont des enregistrements générés artificiellement, imitant les propriétés statistiques de données réelles. Aucun enregistrement d'un jeu de données synthétiques ne correspond à une personne réelle.

ANNEXES

Le cadre d'évaluation s'appuie sur les décisions des juridictions de l'Union résumées ci-dessous. Nous examinons les décisions du Tribunal de l'Union européenne et de la Cour de justice de l'Union européenne (CJUE), y compris les arrêts rendus sur pourvoi. Les décisions des juridictions nationales et des autorités de contrôle sortent du champ de cette annexe. La partie A porte sur le contentieux CRU devant la CJUE. La partie B passe en revue la jurisprudence de la CJUE sur le caractère identifiable qui l'a précédé, de la plus ancienne à la plus récente.

A : LE CONTENTIEUX CRU DEVANT LA CJUE

A1 : CONSEIL DE RÉSOLUTION UNIQUE (CRU) CONTRE CONTRÔLEUR EUROPÉEN DE LA PROTECTION DES DONNÉES

Affaire T-557/20, Conseil de résolution unique contre Contrôleur européen de la protection des données, EU:T:2023:219.

Le Tribunal s'est orienté vers une approche relative du caractère identifiable. Le partage de données pseudonymes avec un tiers, tel que Deloitte, sans transférer également les informations de réidentification, ne rend pas automatiquement les données anonymes du point de vue de ce tiers. Le caractère personnel des données dépend de la question de savoir si le destinataire dispose de moyens légaux susceptibles, en pratique, de lui permettre d'accéder aux informations supplémentaires nécessaires à la réidentification. Le Tribunal a également jugé que des données pseudonymisées ne constituent pas automatiquement des données personnelles dans tous les cas et pour toute personne.

La CJUE a annulé cet arrêt sur pourvoi dans l'affaire C-413/23 P, et a renvoyé l'affaire devant le Tribunal.

A2 : CONTRÔLEUR EUROPÉEN DE LA PROTECTION DES DONNÉES CONTRE CONSEIL DE RÉSOLUTION UNIQUE

Affaire C-413/23 P, Contrôleur européen de la protection des données contre Conseil de résolution unique, EU:C:2025:645.

Lorsqu'une information consiste en des opinions ou des points de vue personnels, sa nature d'expression de la pensée d'une personne la rattache nécessairement et étroitement à cette personne. Aucune analyse distincte du contenu, de la finalité ou de l'effet n'est requise pour établir qu'une telle information se rapporte à une personne identifiable.

Des données pseudonymisées ne constituent pas automatiquement des données à caractère personnel pour toute personne et en toute circonstance. Une même donnée peut être personnelle pour le responsable du traitement, qui détient la clé de réidentification, et anonyme pour un tiers qui ne détient ni cette clé, ni n'a aucun moyen raisonnable de se la procurer. La question pertinente est de savoir si le destinataire concerné dispose de moyens raisonnablement susceptibles d'être utilisés pour identifier la personne concernée. L'obligation d'informer les personnes concernées au titre de l'article 15, paragraphe 1, point d), du règlement 2018/1725 naît au moment de la collecte des données et s'apprécie du point de vue du responsable du traitement. Le CRU devait donc informer les personnes concernées

avant de transférer les commentaires pseudonymisés à Deloitte, indépendamment du caractère personnel de ces données du point de vue de Deloitte.

A3 : CONCLUSIONS DE L'AVOCAT GÉNÉRAL

Affaire C-413/23 P, conclusions de l'avocat général Spielmann, Contrôleur européen de la protection des données contre Conseil de résolution unique, EU:C:2025:59.

L'avocat général a invité la Cour à annuler l'arrêt du Tribunal. Il a confirmé que l'approche relative du caractère identifiable s'applique et que le caractère identifiable dépend des moyens raisonnablement à la disposition du destinataire concerné, et non des capacités théoriques d'un acteur disposant de ressources illimitées.

B : JURISPRUDENCE DE LA CJUE

B1 : SCARLET EXTENDED SA CONTRE SOCIÉTÉ BELGE DES AUTEURS, COMPOSITEURS ET ÉDITEURS SCRL (SABAM)

Affaire C-70/10, Scarlet Extended SA contre Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), EU:C:2011:771.

Les adresses IP collectées par les fournisseurs d'accès à internet constituent des données personnelles, car elles permettent d'identifier les utilisateurs. Cet arrêt a constitué un jalon précoce de la jurisprudence de l'UE établissant que les identifiants techniques de réseau relèvent de la définition de l'article 4.

B2 : PATRICK BREYER CONTRE BUNDESREPUBLIK DEUTSCHLAND

Affaire C-582/14, Patrick Breyer contre Bundesrepublik Deutschland, EU:C:2016:779.

Un fournisseur de services de médias en ligne détient une adresse IP dynamique en tant que donnée personnelle lorsqu'il dispose de moyens légaux pour obtenir, auprès du fournisseur d'accès à internet, des informations supplémentaires d'identification. Ce raisonnement va plus loin encore. Des données par ailleurs impersonnelles peuvent demeurer des données personnelles pour un responsable du traitement disposant de moyens légaux pour accéder à des informations identifiantes détenues par un tiers. Le seul fait que cette information se trouve chez une autre partie ne place pas la personne concernée hors d'atteinte du RGPD.

B3 : PETER NOWAK CONTRE DATA PROTECTION COMMISSIONER

Affaire C-434/16, Peter Nowak contre Data Protection Commissioner, EU:C:2017:994.

L'expression « toute information » figurant à l'article 4 du RGPD traduit un choix législatif délibéré d'attribuer à la notion de donnée personnelle un champ d'application large. Une information se rapporte à une personne identifiée ou identifiable lorsque son contenu, sa finalité ou son effet la rattache à cette personne. Les opinions personnelles, en tant qu'expressions de la pensée d'une personne, renvoient nécessairement à leur auteur. Une analyse fondée sur le seul contenu peut donc suffire. Les trois critères sont des alternatives reliées par « ou », non un test cumulatif.

B4 : GESAMTVERBAND AUTOTEILE-HANDEL E.V. CONTRE SCANIA CV AB

Affaire C-319/22, Gesamtverband Autoteile-Handel eV contre Scania CV AB, EU:C:2023:837.

Un numéro d'identification de véhicule pris isolément n'est pas une donnée personnelle. Il le devient lorsque la partie qui le reçoit dispose des moyens et des informations permettant d'identifier un propriétaire de véhicule qui est une personne physique. Des données par ailleurs impersonnelles peuvent devenir personnelles. L'élément déclencheur est de savoir si le responsable du traitement les transmet à des personnes disposant de moyens raisonnablement susceptibles de permettre l'identification. Lorsque tel est le cas, les données sont personnelles à la fois pour ces personnes et, indirectement, pour le responsable du traitement d'origine.

B5 : OC CONTRE COMMISSION EUROPÉENNE

Affaire C-479/22 P, OC contre Commission européenne, EU:C:2024:215.

L'identification n'est pas raisonnablement probable lorsque la loi l'interdit ou lorsque l'effort en temps, en coût et en moyens serait disproportionné. La perspective qui gouverne l'appréciation du caractère identifiable dépend des circonstances du traitement en cause. La juridiction ne se limite pas à examiner si le responsable du traitement détient des informations identifiantes. Elle doit examiner si le public concerné pourrait raisonnablement identifier une personne à partir des déclarations en cause, y compris en les combinant avec des informations issues d'autres sources.

B6 : IAB EUROPE CONTRE GEGEVENSBESCHERMINGSAUTORITEIT

Affaire C-604/22, IAB Europe contre Gegevensbeschermingsautoriteit, EU:C:2024:214.

Une chaîne de transparence et de consentement constitue une donnée à caractère personnel au sens du RGPD, en particulier lorsqu'elle est associée à une adresse IP. L'arrêt confirme le critère du « contenu, de la finalité ou de l'effet » établi dans l'arrêt Nowak et étend le raisonnement de l'arrêt Breyer à ce contexte. Le point central demeure le même. Lorsque le responsable du traitement dispose de moyens légaux pour obtenir des informations identifiantes auprès d'une autre partie, des données par ailleurs impersonnelles se rattachent néanmoins à une personne identifiable. Peu importe que le responsable du traitement détienne cette information directement ou l'obtienne par l'intermédiaire d'un tiers.

C : DÉFINITIONS ET CONSIDÉRANTS PERTINENTS DU RGPD

C1 : CONSIDÉRANT 26 DU RGPD

(26) Il y a lieu d'appliquer les principes relatifs à la protection des données à toute information concernant une personne physique identifiée ou identifiable. Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable. Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne

physique directement ou indirectement, tels que l'individualisation. Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci. Il n'y a dès lors pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche.

C2 : ARTICLE 4, POINT 1, DU RGPD – DONNÉES À CARACTÈRE PERSONNEL

1) «données à caractère personnel», toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée»); est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ;

C3 : ARTICLE 4, POINT 5, DU RGPD – PSEUDONYMISATION

5) «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 ;

C4 : ARTICLE 25 DU RGPD – PROTECTION DES DONNÉES DÈS LA CONCEPTION ET PROTECTION DES DONNEES PAR DÉFAUT

1. Compte tenu de l'état des connaissances, des coûts de mise en œuvre et de la nature, de la portée, du contexte et des finalités du traitement ainsi que des risques, dont le degré de probabilité et de gravité varie, que présente le traitement pour les droits et libertés des personnes physiques, le responsable du traitement met en œuvre, tant au moment de la détermination des moyens du traitement qu'au moment du traitement lui-même, des mesures techniques et organisationnelles appropriées, telles que la pseudonymisation, qui sont destinées à mettre en œuvre les principes relatifs à la protection des données, par exemple la minimisation des données, de façon effective et à assortir le traitement des garanties nécessaires afin de répondre aux exigences du présent règlement et de protéger les droits de la personne concernée.

2. Le responsable du traitement met en œuvre les mesures techniques et organisationnelles appropriées pour garantir que, par défaut, seules les données à caractère personnel qui sont nécessaires au regard de chaque finalité spécifique du traitement sont traitées. Cela

s'applique à la quantité de données à caractère personnel collectées, à l'étendue de leur traitement, à leur durée de conservation et à leur accessibilité. En particulier, ces mesures garantissent que, par défaut, les données à caractère personnel ne sont pas rendues accessibles à un nombre indéterminé de personnes physiques sans l'intervention de la personne physique concernée.

3. Un mécanisme de certification approuvé en vertu de l'article 42 peut servir d'élément pour démontrer le respect des exigences énoncées aux paragraphes 1 et 2 du présent article.

SAMMAN

LAW & CORPORATE AFFAIRS

privacy response team

Blaeu