(Temps de lecture du document estimé à 9mn)

La data d'intérêt général et l'intérêt général de la data

Nous sommes confrontés à une crise mondiale.

Les décisions que les dirigeants prendront au cours des prochaines semaines façonneront le monde pour les années à venir. L'une des bases de la prise de ces décisions est la disponibilité des bonnes données. Dans la lutte contre le coronavirus, un aperçu des actions préventives, de la mobilité des populations, de la propagation de la maladie et de la résilience des personnes et des systèmes pour faire face au virus, peut aider les responsables de la santé publique et humanitaire à réagir plus efficacement à la pandémie de Covid-19.

Pourtant, aujourd'hui, les leaders de la santé publique qui font des choix difficiles manquent de données de qualité afin de répondre à des questions clés, telles que : où la maladie risque-t-elle de se propager ? Y-a-t-il des domaines prioritaires à contenir pour limiter la propagation ? Où sont les communautés les plus vulnérables ? …

Données et application de tracking : comment cela marche ?

Les applications existantes aujourd'hui (*)

De nombreuses applications mobiles se sont développées intégrant des fonctionnalités de géolocalisation à des finalités de cartographie, mobilité, services géolocalisés, rencontre géolocalisée, … Elles reposent essentiellement sur les technologies de type GPS positionnant un terminal par satellites -à condition de recevoir un réseau- avec une précision au mieux de 10 mètres. La technologie Bluetooth, avec une portée d'environ 10 mètres pour le protocole embarqué dans les smartphones, ne nécessite pas de réseau, mais les terminaux doivent se reconnaître. Il existe d'autres technologies, mais non applicables au sujet actuel.

En dehors de tout débat éthique lié à la réexploitation des données personnelles, une telle application peut significativement contribuer à alerter une personne ayant été à proximité d'un individu infecté. Il y a cependant de multiples risques de "faux positifs" (caissière derrière une vitre protectrice, voisin habitant à l'étage du dessus, ...) ou de "faux négatifs" (batterie vide, difficulté de reconnaissance des terminaux), ou d'imprécisions (a-t-on été en contact à 2 mètres, ou à 15 mètres ?).

Afin de dissiper tout doute autour de l'usage des données qui seront collectées, et d'éventuelles failles de sécurité permettant l'intrusion dans le terminal, on peut recommander la mise à disposition en open-source de 100 % du code de l'application (ce qui sera visiblement partiellement le cas pour l'application STOP COVID proposée par le gouvernement français https://www.lunabee.studio/press/PR_STOPCOVID_26042020_FR.pdf).

(*) Il est difficile d'écrire sur un projet d'application mobile dont on ne connaît pas encore les fonctionnalités, les technologies utilisées, les types d'identifiants utilisés, leur lieu de stockage (local-central, …), l'impact sur la vie de la batterie, …

Les alternatives

La réponse à la crise sanitaire est constituée d'un ensemble d'actions prises dans différents domaines tels que la distanciation sociale (confinement), les aides financières (report de remboursement de prêt, étalement de charges sociales, …), les aménagements du travail (home office, chômage partiel), l'épidémiologie (recherche de vaccins, …).

Parmi ces domaines, il en est un en particulier, la détection des porteurs et leur suivi, pour lequel la donnée a un véritable rôle à jouer. Le tracking des individus est une solution envisagée pour répondre à cette problématique.

Face au débat sociétal soulevé par la création d'"applications pour limiter la contagion", on retiendra que le problème posé est "**limiter la contagion**". Ci-dessous des éléments factuels pour éviter d'assimiler le problème **limiter la contagion**, la solution **le tracking des individus** et le contexte **la privauté des données personnelles.**

Premièrement, il existe des alternatives à la technique de tracking basé sur les traces GPS ou sur les échanges Bluetooth, notamment le QR Code. Avec les QR Code, le tracking devient alors déclaratif citoyen et peut se limiter à la partie la plus sensible des activités telles que les déplacements. Face à l'argument de l'inefficacité potentielle des systèmes déclaratifs, il faut rappeler que les tracking imposés font l'objet de nombreux biais notamment de précision.

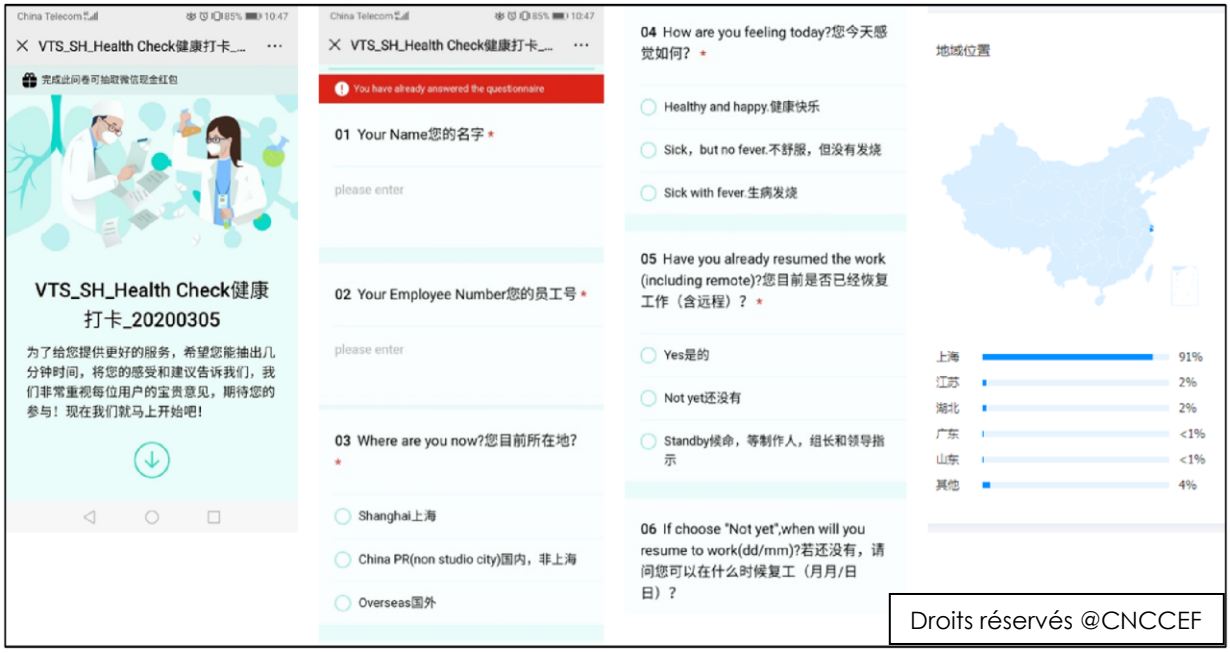
Si on exclut la question sociétale du respect des données privées, l'efficacité du tracking comme réponse à la crise sanitaire peut être questionné :

- Précision des données (GPS et Bluetooth peuvent confondre des individus si trop de proximité).
- Exhaustivité du suivi : quelle que soit la forme du traçage, on restera toujours sur des outils d'estimation fournissant des informations à titre indicatif.
- S'agissant d'un suivi déclaratif, surgissent d'autres problèmes de précision : tout le monde ne déclare pas, il peut également y avoir de la mauvaise fois, plus simplement encore : une partie de la population n'a pas accès à la déclaration (en France, 1 téléphone mobile sur 5 n'est pas un smartphone et 5 % de la population n'est pas équipée).

Deuxièmement, une application est uniquement un moyen et il existe des alternatives pour exploiter les données comme le montre la liste d'innovations suivantes lancées en Chine avec la crise du Covid-19 (source : rapport CCE Chine 20 mars 2020 - COVID19 - Accélérateur de l'innovation).

- Identification des malades et traçabilité des individus : questionnaire santé sous forme de mini application mobile.

Exemple de solution déclarative à l'échelle d'une entreprise pour soutenir l'organisation dans la reprise de l'activité :



- Les géants du Net mettent à disposition les données qu'ils utilisent déjà pour le tracking commercial (Google + Apple) et les opérateurs télécom mettent à disposition les historiques de déplacement.
- Diverses initiatives liées au QR Code (déclaratif) dans les transports en commun. Le QR Code est un moyen estimatif basé sur la déclaration et peut être rendu incitatif pour avoir accès à certains services (accès à des restaurants et magasins par exemple).
- Attestation de mobilité par QR Code pour les déplacements en voiture pour démontrer l'activité des 14 derniers jours. Support au déconfinement.
- Livraison sans contact (véhicules autonomes et drones) :

- In situ : reconnaissance faciale (port du masque OK) et contrôle en masse de la température corporelle.
- Accélérer la recherche scientifique et trouver plus rapidement les remèdes par le drug screening qui est la combinatoire massive des composants médicamenteux.
- Détection / diagnostic rapide des cas - SAN10 : Alibaba diagnostic de Covid par scan CT (96 %).
- SER01 Information du public : carte dynamique des cas + infos temps réel.
- Tencent : applications anti-rumeur.
- Aide au contrôle des forces de l'ordre et des équipes de contrôle sanitaire.

- En Israël, le tracking était déjà en place pour lutter contre le terrorisme, de ce fait la population s'en remet à l'armée et aux services secrets.
- Le bracelet électronique pour les anciens malades est une technologie éprouvée car déjà en place pour le contrôle des libertés surveillées.

Les biais récurrents de la donnée

Covid-19 permet à 100 % de la population d'être confrontée à des analyses de données en continu pour se faire son propre jugement. Comme pour des arguments en politique, on peut faire dire à une même source de données tout et son contraire. Pourtant il existe une acceptation sur le fait que les analyses quantitatives sont "vraies" puisqu'avec des données on est capable de justifier un OUI ou un NON. Ainsi les données des décès du Covid-19 devraient être fiables. Or nous le constatons, les chiffres font débat. L'occasion de revenir sur un aspect méconnu des analyses quantitatives que sont les biais. Voici quelques exemples :

1- Les données ne sont pas immédiatement disponibles : au début de la crise le nombre de décès journaliers en France ne comprenait pas les décès en EHPAD, pourtant c'est bien ce chiffre qui était utilisé pour se comparer aux autres pays, conduisant à une sous-évaluation. Les chiffres ne contiennent d'ailleurs toujours pas les données de personnes décédées à leur domicile.

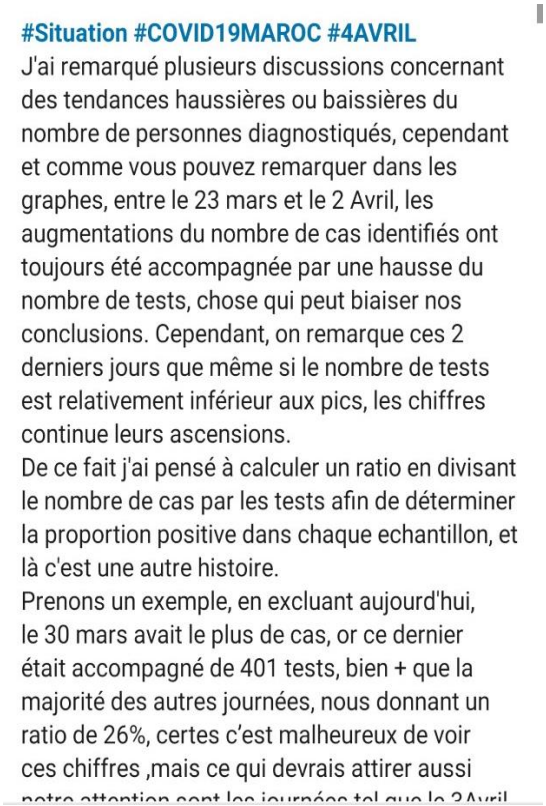
2- Les données sont imprécises : c'est le cas des données des capteurs en général. Il suffit de comparer 2 sondes thermostatiques pour s'en convaincre ou encore de lire les articles sur la précision d'une trace GPS/Bluetooth.

3- Les données sont fausses. En effet l'absence de tests sur une pathologie présentant un diagnostic différentiel de cette taille rend très fortement probable l'inclusion de personnes contaminées ou décédées du fait d'autres pathologies présentant des symptômes similaires.

4- Les données ne sont pas homogènes : chacun les a collectées selon des méthodes différentes. Par exemple, le nombre de cas détectés est biaisé dans la mesure où certains pays utilisent des tests sur l'ensemble de la population, tandis que d'autres ne testent que les patients qui se déclarent. Cette absence d'homogénéité entraîne une impossibilité d'utiliser les données d'autres pays à des fins d'extrapolation ou d'apprentissage. Le site https://coronavirus.jhu.edu/map.html montre très clairement que, sous l'hypothèse d'avoir une létalité et une contagiosité homogène, les données sont clairement hétérogènes. Ainsi les comparaisons entre pays ne font aucun sens.

5- La mauvaise foi ou la manipulation : il est difficile de s'en remettre à des gouvernements qui jouent leur popularité et donc peuvent être à même de maquiller les chiffres. Dans le cas du Covid-19, le doute persiste sur la Chine dont le nombre de morts est particulièrement faible par rapport à des nations plus petites avec une densité de population plus faible.

6- La mauvaise maîtrise des règles statistiques (cf. exemple ci-dessous). On parle aussi de redressement des données en statistiques. En effet, les conclusions sont tirées sur des données redressées de façon erratique, des échantillons non stables et des redressements statistiques homogènes sur des données dont les stratégies de collectes n'ont fait qu'évoluer.



#Situation #COVID19MAROC #4AVRIL

J'ai remarqué plusieurs discussions concernant des tendances haussières ou baissières du nombre de personnes diagnostiqués, cependant et comme vous pouvez remarquer dans les graphes, entre le 23 mars et le 2 Avril, les augmentations du nombre de cas identifiés ont toujours été accompagnée par une hausse du nombre de tests, chose qui peut biaiser nos conclusions. Cependant, on remarque ces 2 derniers jours que même si le nombre de tests est relativement inférieur aux pics, les chiffres continue leurs ascensions.

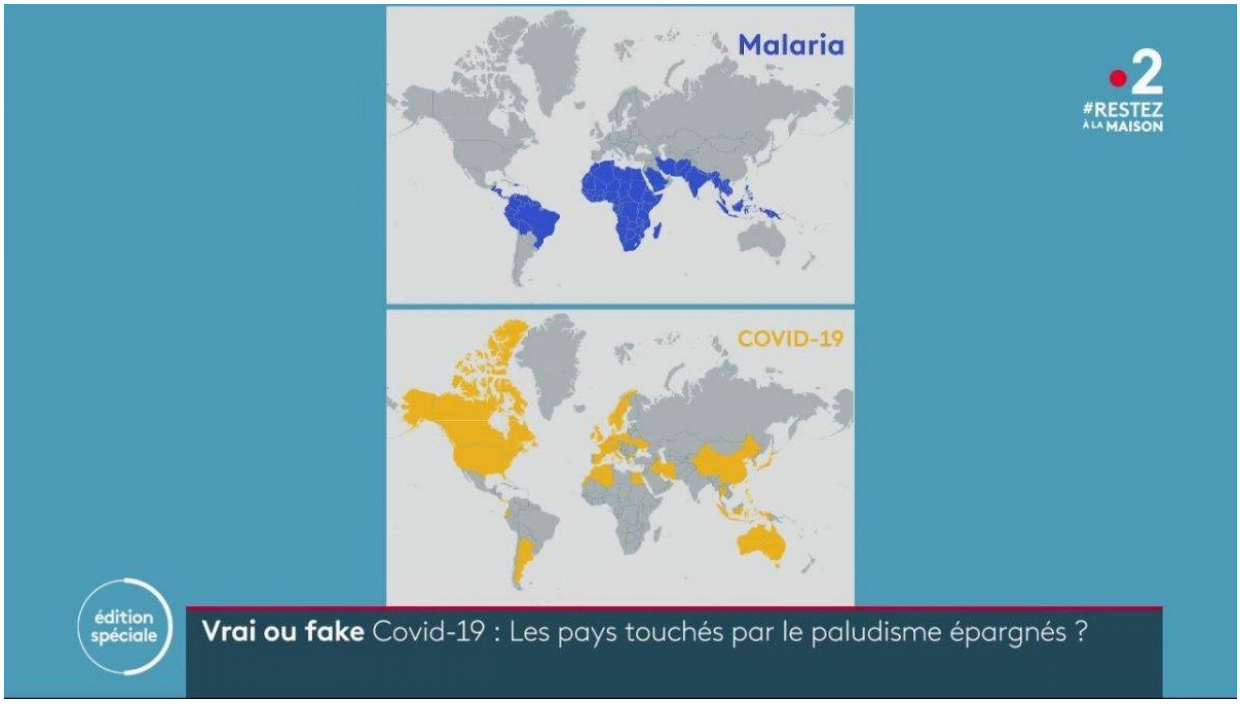
De ce fait j'ai pensé à calculer un ratio en divisant le nombre de cas par les tests afin de déterminer la proportion positive dans chaque echantillon, et là c'est une autre histoire.

Prenons un exemple, en excluant aujourd'hui, le 30 mars avait le plus de cas, or ce dernier était accompagné de 401 tests, bien + que la majorité des autres journées, nous donnant un ratio de 26%, certes c'est malheureux de voir ces chiffres ,mais ce qui devrais attirer aussi

7- Le processus de collecte varie dans le temps : les analyses reposent sur des transactions, or si les systèmes transactionnels changent, les paramètres qui sont liés peuvent varier, les référentiels peuvent être modifiés aboutissant à un classement différent des transactions qui ne peut être détecté au niveau de l'analyse de masse.

8- Les habitudes changent soudainement : il ne s'agit pas véritablement d'un biais, mais comme pour les biais les changements brutaux mettent fin à la capacité d'un algorithme à prédire correctement étant un système de détection des corrélations observées par le passé.

9- Les biais cognitifs sont aussi nombreux : on parle de biais de confirmation lorsqu'on fait dire à la donnée ce qu'on veut qu'elle dise (cf. exemple ci-dessous : cartographie des populations atteintes de palud qui ne sont pas atteintes par le Covid-19), ou d'illusion de concentration induite actuellement par les médias dont 90 % du contenu est porté sur le Covid-19. Le système 1 du cerveau (cf. Daniel Kanheman) et son analyse instantanée prend le pas sur le système 2, base de la réflexion construite. En d'autres termes, dans une période aussi anxiogène les biais sont naturels, les réactions émotionnelles prennent le pas sur les réactions rationnelles.



Et Demain ? Après ?

Les données : qualité ? fraîcheur ?

Les éléments de qualité et de fraîcheur des données ne sont pas à considérer dans l'absolu, mais doivent être considérées relativement à ce que l'on veut en faire. Compte tenu des biais énumérés précédemment, le plus important sera, si ce n'est de tenter de les corriger, de ne pas les omettre avant de tirer des conclusions qui pourraient s'avérer hâtives.

En effet, les données Covid-19 aujourd'hui ne sont pas remontées en temps réel, elles ne le sont qu'une fois que les personnes se sont présentées à l'hôpital avec les symptômes, ou qu'elles en sont décédées pour ce qui concerne les EHPADs. Pour un confinement localisé et efficace, une remontée de données plus rapide est nécessaire, les canaux restant à définir (application, sms, etc.) pour se déclarer atteint, avec toutes les précautions éthiques qu'il faudra prendre pour éviter les débordements. Ces canaux pourraient à terme être généralisés pour permettre des remontées rapides d'informations (non nécessairement liées au Covid-19 une fois la crise passée) offrant une possibilité de traitement "at the edge" et donc une réaction rapide des autorités en cas de problèmes. Car l'actionnabilité de la donnée vient en grande partie de notre capacité à la traiter rapidement.

Il faudra aussi fiabiliser les données le plus possible par des tests (ex ante) ou des contrôles (ex post), pour éviter de ne biaiser les tentatives de modélisation et de déclencher des actions inappropriées. Si aujourd'hui les données sont de qualité suffisante pour évaluer une tendance, cette qualité n'est pas suffisante pour pouvoir déclencher des actions à un niveau plus granulaire. En d'autres termes, et pour reprendre l'exemple de la situation actuelle, le tamis est aujourd'hui trop important. Une fois le confinement terminé, considérant le taux de contamination et le nombre de personnes ayant déjà été atteintes, nous repartirions vers un schéma de crise identique à celui initialement vécu.

Une acculturation de la population aux biais que comportent les données est nécessaire pour éviter une incompréhension des actions menées par les gouvernements. En effet, aujourd'hui, les données remontées -surtout lorsque celles-ci sont quantifiées- apparaissent comme un élément de vérité, or elles ne sont qu'une approximation. Sans cette acculturation, aucun des biais cognitifs ne sera diminué.

Partage ? Données personnelles ?

Le RGPD est parfois utilisé comme argument pour s'opposer à des dispositifs de partage, de collecte et de traitement de données personnelles, y compris pour les données relatives à la pandémie de Covid-19.

Il est utile de rappeler que, dans la grande majorité des cas, le RGPD n'interdit pas le traitement de données personnelles, mais l'encadre par diverses mesures visant à protéger l'intérêt des personnes concernées. Il convient donc plutôt de regarder l'impact de la réglementation, et notamment du RGPD, sur les conditions de mise en œuvre de ces traitements, et plus particulièrement sur ceux liés à une application de tracking.

En premier lieu un traitement, pour être licite, doit disposer d'une base légale définie à l'article 6 du RGPD. La CNIL a rappelé, lors de l'envoi du SMS d'alerte sur la mise en place du confinement, que le règlement général sur la protection des données (RGPD) permet l'utilisation de données personnelles sans consentement des personnes, notamment dans le cadre d'une obligation légale, de missions d'intérêt public ou pour la sauvegarde des intérêts vitaux des personnes (alinéa d et e de l'article 6 du RGPD). Un tel argument pouvant être légitimement soumis à débat sur son interprétation, la mise en œuvre d'une application de tracking s'appuiera plutôt sur le volontariat des participants, c'est-à-dire en faisant appel au consentement explicite et éclairé des personnes concernées. Eclairé signifie notamment que les personnes doivent être précisément informées lors de la sollicitation de leur consentement de la finalité principale et des finalités secondaires éventuelles du traitement, de la nature des données concernées par le traitement, des transferts éventuels de ces données et des conditions de conservation et de destruction des données. En l'occurrence, pour l'application de tracking envisagée, la finalité principale est de pouvoir identifier les personnes ayant eu un contact probable avec une personne contaminée et de les en informer pour les inciter à se mettre en quarantaine :

- La donnée relative aux personnes contaminées ou au cas contact, identifiée par Bluetooth et non géolocalisée, sera a priori anonyme et ne pourra donc être transmises à aucun autre tiers, sauf si cela était explicitement signalé dans les finalités du traitement.
- Les principes de limitation du RGPD conduisent à n'autoriser l'accès à ces données qu'aux seules personnes en ayant une utilité légitime, c'est-à-dire a priori aux seules personnes en charge de la maintenance de l'application.
- La conservation de la donnée devant correspondre à la durée utile au traitement, il est raisonnable d'estimer qu'elle ne peut excéder la durée de quarantaine de 14 jours préconisée pour la pandémie de Covid-19. Les données de signalement éventuellement recueillies devraient donc être effacées au terme d'une période de 14 jours, ou anonymisées pour permettre la réalisation éventuelle de travaux de nature statistique par agrégation.

Par ailleurs, eu égard notamment à la volumétrie du nombre de personnes concernées et au risque élevé présenté pour les droits et libertés des personnes concernées, un projet d'application de tracking fera obligatoirement l'objet d'une analyse d'impact préalable soumise à la validation de la CNIL, autorité de contrôle. Cette analyse d'impact doit notamment présenter de manière détaillée l'ensemble des mesures techniques et organisationnelles prises pour assurer la sécurité des données, présentation soumise à la validation d'experts sur les différents risques relatifs aux données personnelles, et en particulier ceux liés à la cybersécurité.

Le RGPD permet donc de définir un cadre permettant de garantir le "bon usage" de la donnée partagée dans le cadre d'une application de tracking Covid-19, ainsi que la conception technique associée. Il reste donc simplement à veiller à son application stricte et rigoureuse, même en période de crise sanitaire.

Contrôle ? Jusqu'où ira-t-on ?

La façon dont les gouvernements utilisent les données pour réagir à la crise du Covid-19 est de plus en plus préoccupante. Alors que de nouvelles technologies émergent qui visent à recueillir, diffuser et utiliser des données afin de soutenir la lutte contre le Covid-19, nous devons nous assurer qu'elles respectent les meilleures pratiques éthiques. Même en temps de crise, nous devons nous conformer aux règles de confidentialité des données et nous assurer qu'elles sont utilisées de manière éthique. Sachant bien que l'éthique et la morale sont bien différents en fonction des cultures et des mœurs.

L'une des façons de le faire est d'établir des comités éthiques indépendants. Leur rôle sera de créer des mécanismes de gouvernance des données pour trouver l'équilibre entre les intérêts publics concurrents, tout en protégeant la vie privée individuelle. Parmi les exemples de ces règles, mentionnons la mise en place de lignes directrices claires sur l'objet et le calendrier de l'utilisation des données, la définition de processus clairs pour l'accès, le traitement et la résiliation des données personnelles à la fin de la crise.

Malheureusement ces bonnes pratiques sont souvent longues à mettre en place, et difficilement compatibles avec la situation actuelle qui nous pousse à prendre des mesures dans l'urgence.

Tedros Adhanom Ghebreyesus, directeur général de l'Organisation Mondiale de la Santé, a déclaré : "Vous ne pouvez pas combattre un feu les yeux bandés". Les bonnes informations entre les mains des bonnes personnes peuvent sauver des vies en temps de crise. Il sera essentiel de veiller à ce que de telles mesures de surveillance de la santé ne prévalent pas au-delà des circonstances extrêmes auxquelles nous sommes confrontés aujourd'hui, afin que les gens n'aient pas l'impression de perdre leur vie privée dans un nouvel ordre mondial.

En conclusion

Les phénomènes d'ampleurs produisent des réactions de toutes natures, politiques, sociales, économiques, psychologiques, … Toutes ces réactions pressent les dirigeants d'apporter une réponse à l'inquiétude générée à différents niveaux.

La data informe les populations et les gouvernements et les invite à un processus décisionnel. Chacun de nous a suivi l'évolution des données lorsque le foyer se trouvait à Wuhan, mais le risque est filtré par nos émotions, nos expériences et nous empêche de saisir individuellement l'ampleur du phénomène.

Pourtant, un accès à des données démocratisées, voire vulgarisées, devrait devenir la norme afin de modéliser la perception du risque pour préparer chacun de nous à affronter la situation et à éviter les déviances que nous avons pu observer : incertitudes du corps médical sur le traitement à apporter, défiance des citoyens sur les gestes et comportements à adopter, désaccords des dirigeants sur les décisions à prendre. Quand nous sommes au cœur, la place n'est plus au débat mais aux remèdes. La data, avec sa fraîcheur et qualité adéquates, devient alors l'indicateur de la situation pour laisser aux experts le soin de procéder aux actions nécessaires.

La data n'a pas pour ambition de traiter les malades, quoique, si on regarde de près, elle pourrait aider. Toutefois, elle peut être utile dans la gestion de la pandémie, de l'épidémie, et apporter un éclairage dans le traitement de la maladie.

La data d'intérêt général et l'intérêt général de la data.

Auteurs

Jean-David Benassouli (PwC), Marc Bidou (Bilendi), Jonas Cadillon (Roland Berger), Arnold Haine (BVA Group), Bertrand Hassani (Deloitte), Benoit Heitz (Onepoint), Hichem Lazrek (Menway Carrières), Pascal Lefort (Strategir), Olivier Leroy (PMP), Hervé Tranger (BVA Group)