La gouvernance des données d'origine humaine

Par Teresa Scassa

Les plateformes collectent de grandes quantités de données sur les utilisateur·rice·s à des fins commerciales et opérationnelles diverses, notamment pour mettre en place des programmes de publicité ciblée. Ces données peuvent également être partagées avec des tiers. Les données des plateformes sont également grattées par divers acteur·rice·s, notamment des chercheur·euse·s, des acteur·rice·s de la société civile, des concurrents commerciaux et des courtier·ère·s en données. Un exemple notoire de grattage de renseignements personnels à partir de plateformes est celui de Clearview AI, qui a gratté des photographies en ligne pour bâtir sa gigantesque base de données de reconnaissance faciale.

Les données les plus recherchées par les plateformes sont généralement celles qui proviennent des humains et de leurs activités. Les questions relatives à la légitimité de la collecte, de l'utilisation ou du partage de ces données sont souvent liées à la question de savoir si les données concernent des personnes identifiables ou si elles sont dépersonnalisées à la source, ou encore anonymisées par la suite. Cet article fait valoir que, pour traiter les questions relatives à la collecte et à l'utilisation appropriées des données (sur les plateformes et ailleurs), il ne suffit plus de se demander si ces données concernent une personne identifiable (c'est-à-dire s'il s'agit de données à caractère personnel) ou si elles sont anonymisées. En effet, si la différence entre « données personnelles » et « données anonymes » reste une distinction importante dans le cadre de la législation sur la protection des données, il est toutefois nécessaire de définir le concept de « données d'origine humaine » dans un cadre de gouvernance distinct.

À l'heure actuelle, les lois sur la protection des données (et la protection de la vie privée) régissent généralement les renseignements relatifs à des *personnes identifiables*. En effet, ces lois reposent sur des principes de confidentialité qui protègent le droit des individus à contrôler les renseignements les concernant. Dans un tel cadre, si les renseignements en question ne peuvent être reliés à une personne en particulier, le traitement de ces derniers ne peut avoir d'incidence sur le droit à la protection de la vie privée de cette personne. Par conséquent, dans un contexte où les données sont essentielles pour l'innovation basée sur les données, y compris l'intelligence artificielle (IA), il n'est pas surprenant qu'il y ait une pression considérable pour faire la distinction entre les données personnelles et les données anonymisées, et pour placer ces dernières en dehors du champ d'application des lois sur la protection des données.

En effet, le fait d'établir une distinction entre les données personnelles et les données anonymisées permet une utilisation plus large de ces dernières. Telle est la position normative adoptée dans le *Règlement général sur la protection des données* (RGPD) de l'Union européenne. Cette position apparaît également de façon évidente dans la partie du projet de loi C-27 relatif à la *Loi sur la protection de la vie privée des consommateurs* et dans des lois telles que la *Loi sur la protection des renseignements personnels sur la santé* de l'Ontario. En règle générale, les données personnelles font l'objet d'une réglementation et d'une gouvernance, tandis que les données anonymisées peuvent être utilisées à l'insu des personnes concernées et sans leur consentement. Tout contrôle de l'utilisation de données anonymisées porte généralement sur le processus

d'anonymisation (car, si l'anonymisation n'est pas effectuée correctement, les données peuvent être reliées à une personne identifiable et donc rester des données personnelles). De plus en plus, les lois sur la protection des données prévoient également des sanctions en cas de réidentification délibérée de données anonymisées.

Cette approche relative aux données anonymisées pose différents problèmes. Tout d'abord, un nombre croissant d'universitaires et de défenseur·euse·s de la vie privée ont averti que, compte tenu des volumes actuels de données et d'outils d'analyse de données, il sera toujours possible de réidentifier des individus à partir d'ensembles de données, ce qui fait de l'anonymisation une chimère. Cela dit, ce point ne constitue pas nécessairement un problème en vertu des lois sur la protection des données, car celles-ci définissent souvent l'anonymisation en termes relatifs. Ainsi, la question de la réidentification nécessite la prise en compte de divers facteurs, notamment la sensibilité des données, les autres données pertinentes disponibles qui pourraient contribuer à la réidentification des personnes et la probabilité qu'un·e adversaire cherche à réidentifier une ou plusieurs personnes à partir de l'ensemble de données anonymisées. L'une des difficultés réside dans le fait que le risque de réidentification pour les ensembles de données anonymisées peut évoluer au fil du temps, à mesure que de plus en plus de données deviennent disponibles et que de nouveaux outils analytiques sont mis au point.

De plus, certain·e·s chercheur·euse·s soutiennent qu'il serait utile de définir un concept portant sur la protection de la vie privée des groupes ou collectives et qui reconnaisse les potentiels intérêts de groupe associés aux données humaines collectées, et ce même si ces données sont anonymisées. Toutefois, les lois actuelles sur la protection des données ne reconnaissent pas la vie privée des groupes, même si ces arguments ont été repris non seulement par les universitaires, mais aussi par les défenseur·euse·s de la vie privée dans de nombreux contextes. Les débats sur la gouvernance de l'IA soulèvent également des préoccupations quant au fait que même les données anonymisées peuvent avoir des répercussions négatives sur les individus et/ou les groupes. Si elle est approuvée, la partie du projet de loi C-27 portant sur la *Loi sur l'intelligence artificielle et les données*, par exemple, engendrerait des obligations en matière de gestion des données anonymisées, en particulier en ce qui concerne leur potentiel à conduire les IA à fournir des résultats biaisés.

La distinction entre les données personnelles et les données anonymisées laisse subsister une importante lacune en matière de gouvernance. Pour combler cette lacune, nous devons reconnaître une nouvelle catégorie de données que j'appelle « données d'origine humaine ». Les données d'origine humaine sont des données dérivées de l'humain ou de ses activités, mais qui ne sont pas des données personnelles. Alors que la base normative du droit à la protection de la vie privée est l'autonomie et la dignité des individus, la base normative de la gouvernance des données d'origine humaine repose sur les droits fondamentaux de la personne. Les individus et/ou les communautés/groupes auxquels ils ou elles appartiennent sont la source de ces données. Or, celles-ci peuvent être utilisées de manière à nuire ou à exploiter le collectif ou ses membres, et elles nécessitent une certaine forme de gouvernance pour garantir qu'elles ne sont pas utilisées de manière préjudiciable, discriminatoire ou limitant les droits des personnes.

Pourquoi la protection de la vie privée n'est-elle pas une base normative suffisante pour la gouvernance des données d'origine humaine? L'une des raisons à cela est que le droit relatif à la

protection de la vie privée se fonde sur l'individu et son droit à contrôler ses données personnelles et, par le biais de ses données, son identité. Le concept axé sur la protection de la vie privée des groupes concerne des droits plus collectifs sur les données et permet souvent de réaliser des avancées pour répondre aux préoccupations des groupes en quête d'équité. Étant donné que les données d'origine humaine peuvent avoir des répercussions sur les décisions prises à l'égard des groupes et des individus d'une manière qui va au-delà de la protection de la vie privée, la bonne gouvernance de ce type de données peut contribuer à garantir des droits de la personne autres que la protection de la vie privée, tels que le droit de ne pas être victime de discrimination.

En mettant l'accent sur les données d'origine humaine (plutôt que sur les données personnelles anonymisées), on place l'*humain* et non l'*individu* au cœur de l'analyse. Il s'agit d'une approche plus explicitement fondée sur les droits de la personne. Il ne suffit pas d'adjoindre aux lois existantes le principe de la gouvernance des données personnelles anonymisées. En effet, même si certaines données d'origine humaine sont des données personnelles qui ont été anonymisées par la suite, d'autres sont collectées dans des contextes où elles ne sont jamais liées à des personnes identifiables. Par exemple, les données relatives à la présence du virus de la COVID-19 et de ses variants dans les eaux usées ne sont pas collectées de manière à permettre l'identification de personnes spécifiques. Il ne s'agit donc jamais de données personnelles, mais bien de données d'origine humaine. Les données anonymisées ne sont donc qu'un sous-ensemble au sein de la catégorie des données d'origine humaine.

Parce qu'elle se distingue de l'individu qui se situe au cœur de l'approche normative adoptée par le droit à la vie privée, la gouvernance des données d'origine humaine peut englober un plus grand nombre de considérations, dont des préoccupations plus larges en matière de droits de la personne (comme le droit de ne pas subir de discrimination), ainsi que des principes éthiques. Pour la collecte et l'utilisation de données d'origine humaine, on pourrait exiger la transparence et l'engagement du public. On pourrait également exiger un accès ouvert aux résultats des analyses ou de la recherche, ou le retour d'avantages directs ou indirects à la communauté (plutôt que de supposer que la collecte des données favorisera l'innovation ou l'économie). Ces concepts ont déjà été évoqués dans les discussions sur les données ou les connaissances en tant que biens communs, l'éthique de la science citoyenne, ainsi que les cadres relatifs au partage des avantages liés à l'accès aux données pour l'utilisation des ressources génétiques. Les facteurs utiles pour déterminer la gouvernance appropriée des données d'origine humaine peuvent varier en fonction du contexte. Par exemple, la gouvernance des données d'origine humaine peut tenir compte de la nature ou de la composition des communautés auprès desquelles ces données sont collectées, ainsi que de la relation entre la collecte des données et les infrastructures publiques ou privées.

Les arguments en faveur de la gouvernance des données d'origine humaine ne vont pas nécessairement à l'encontre de la loi sur la protection des données (qui reste nécessaire) ni de la gouvernance des données anonymisées, en particulier en ce qui concerne les protections visant à garantir qu'elles restent anonymes. Il s'agit plutôt d'arguments selon lesquels la collecte omniprésente de données sur les activités humaines et l'utilisation croissante de ces dernières pour prendre des décisions concernant les communautés et les individus dans tous les secteurs et toutes

les sphères d'activité nécessitent un cadre approprié pour garantir la transparence et l'engagement, ainsi que pour protéger les droits de la personne.

Teresa Scassa est professeure et titulaire de la Chaire de recherche du Canada en politiques et droit de l'information à l'Université d'Ottawa. *Elle remercie les participant·e·s à l'atelier de février 2023 sur la gouvernance des plateformes pour les commentaires avisés qu'ils et elles lui ont faits sur une version antérieure de cet article.*