LES MUTATIONS DE LA MISE À DISPOSITION DE CONTENUS AUDIOVISUELS À L'ÈRE DU NUMÉRIQUE : CONSÉQUENCES ET ENJEUX

Rapport 1

Le rôle des données et des algorithmes dans l'accès aux contenus

2017
JANVIER

Le CSA Lab est un groupe de réflexion prospective réunissant des experts du numérique et de l'audiovisuel avec l'objectif d'anticiper et de caractériser les évolutions de l'économie et de la régulation audiovisuelles induites par la transformation numérique. Ses travaux permettent d'enrichir la réflexion du Conseil supérieur de l'audiovisuel, d'éclairer certaines de ses orientations et d'anticiper les évolutions du secteur qui pourraient influer sur ses activités.

Les analyses exprimées par le CSA Lab dans le cadre de ses travaux n'engagent pas le Conseil supérieur de l'audiovisuel et ne peuvent préjuger du sens de ses décisions et avis.

2017

JANVIER

Sommaire

Introduction

I. Le CSA Lab, un groupe de réflexion prospective sur l'audiovisuel

Le secteur audiovisuel est sujet à la transformation numérique que connaît l'économie en général. Dans la grande majorité des secteurs d'activité, de nouveaux acteurs et de nouveaux services ont émergé, modifiant les usages, les équilibres économiques et les chaînes de valeur.

En tant que régulateur du secteur audiovisuel, le CSA porte à ces mutations une attention marquée, aussi bien sur le plan national que sur le plan européen. Il contribue en effet activement à la réflexion de l'ERGA[1], sur la révision de la directive SMA[2], en cours, qui vise notamment à prendre en compte l'émergence des nouveaux acteurs du numérique.

Pour enrichir son regard et pousser plus avant ses analyses, le CSA a souhaité se doter d'un comité de réflexion prospective. Intitulé *CSA Lab* et lancé le 14 juin 2016, ce comité réunit des experts du numérique et de l'audiovisuel dans le but d'anticiper et de caractériser les évolutions de l'économie et de la régulation audiovisuelles induites par la transformation numérique.

- **Ses objectifs et missions**

Le secteur audiovisuel a engagé une transition numérique qui s'intensifie et s'accélère. Il s'affranchit progressivement des frontières géographiques et explore des territoires économiques et technologiques nouveaux, sur lesquels le Conseil a entrepris des travaux depuis plusieurs années. Dans ce contexte, le Lab a pour ambition :

- d'enrichir l'appréhension qu'a le Conseil des enjeux du numérique ;
- d'éclairer certaines orientations du Conseil ;
- d'identifier les évolutions du secteur qui pourraient influer ses activités.

Le Lab n'a pas vocation à se substituer aux travaux du Conseil mais à les compléter d'une manière originale par une approche et une méthodologie différentes.

Le CSA Lab se réunit quatre fois par an et s'est donné pour objectif de publier un ou deux rapports par an. Le présent document constitue son premier rapport.

[1] ERGA : *European Regulators Group for Audiovisual Media Services.*

[2] Directive 2010/13/UE du Parlement européen et du Conseil du 10 mars 2010 visant à la coordination de certaines dispositions législatives, réglementaires et administratives des États membres relatives à la fourniture de services de médias audiovisuels (directive *Services de médias audiovisuels,* dite directive SMA).

- **Sa composition**

Le Lab, présidé par les conseillers Nathalie Sonnac et Nicolas Curien, réunit des personnalités qualifiées aux profils diversifiés, à l'expertise reconnues, d'économie numérique et d'analyse juridique :

- **Maya Bacache-Beauvallet**, professeure de sciences économiques à Telecom ParisTech ;
- **Yann Bonnet**, secrétaire général du Conseil national du numérique ;
- **Olivier Henrard**, rapporteur public au Conseil d'État, ancien secrétaire général du groupe SFR ;
- **Pascale Idoux**, professeure de droit public à l'université de Montpellier ;
- **Winston Maxwell**, associé du cabinet Hogan Lovells, spécialisé dans les médias et les nouvelles technologies ;
- **Francesca Musiani**, chargée de recherche CNRS à l'Institut des sciences de la communication, chercheuse associée à Mines ParisTech ;
- **Marc Tessier**, président de Video Futur Entertainment Group SA, président du Forum des images et membre du Conseil national du numérique.

II. Les thèmes développés par le CSA Lab

Les membres ont souhaité ancrer les premiers travaux prospectifs du Lab aux évolutions envisagées dans le projet de révision de la directive SMA. Ils ont choisi, à ce titre, le thème de la mise à disposition de contenus audiovisuels : quelle évolution de la distribution ? Quelle requalification des distributeurs d'un point de vue juridique et économique ? Quel rôle des données et des algorithmes dans la présentation des œuvres ?

Les travaux ont été décomposés en trois thèmes, traités dans le cadre de trois ateliers parallèles, comprenant chacun trois membres.

- **Le contenu audiovisuel à l'ère numérique**
  (Maya Bacache-Beauvallet, Nathalie Sonnac et Marc Tessier)

L'avènement du numérique a entraîné une mutation du secteur de l'audiovisuel et a fait évoluer en profondeur ses équilibres économiques. Une multitude d'acteurs ont émergé, à l'articulation entre utilisateur et contenu, bousculant les positions établies et la distribution des contenus et occupant d'ores et déjà, pour certains, une place déterminante. Même si plusieurs études tendent à montrer que cette restructuration du secteur pourrait dans une certaine mesure avoir un impact positif sur la qualité et la diversité des contenus, et donc *in fine* sur l'utilisateur, la valorisation des œuvres et le financement de la création restent des questions ouvertes et centrales dans cet écosystème devenu mondial.

- **Qualification juridique des distributeurs et du champ d'application de régulation**
  (Olivier Henrard, Pascale Idoux et Winston Maxwell)

L'évolution des usages, des terminaux et des réseaux a contribué au développement de nouveaux intermédiaires qui bouleversent la distribution des contenus médias. Cet environnement renouvelé pose la question de l'adaptation du statut de distributeur, tel qu'inscrit dans la loi du 30 septembre 1986 relative à la liberté de communication. Il s'agit notamment de mesurer si la régulation audiovisuelle « traditionnelle » doit être adaptée pour englober ces nouveaux acteurs ou, si à l'opposé, ces derniers doivent être appréhendés de manière spécifique. L'analyse a été menée à travers le prisme des objectifs de politique publique poursuivis, et par l'étude, à l'échelle européenne et internationale, de plusieurs cadres réglementaires en matière de distribution et services.

- **Le rôle des données et des algorithmes dans l'accès aux contenus**
  (Yann Bonnet, Nicolas Curien et Francesca Musiani)

La « mise en données » du monde n'épargne pas les médias. Les questions qui découlent de cette transformation sont d'autant plus prégnantes que les données liées à la consommation de contenus audiovisuels comportent une fonction identitaire très forte. Les algorithmes qui traitent ces données sont nombreux et variés : moteurs de recommandation, publicité programmatique, etc. Ils peuvent par exemple fonctionner sur une mise en correspondance (*matching*) ou être fondés sur une approche statistique ou sémantique. Leur influence demande ainsi à être appréhendée selon leur principe de construction et selon la finalité poursuivie. Une typologie des algorithmes utilisés dans le secteur audiovisuel est mise en regard des grands objectifs de politique publique d'un développement numérique durable : diversité culturelle, protection des consommateurs et confiance.

• • •

Le rôle des données et des algorithmes dans l'accès aux contenus

Par définition, une donnée est un élément brut qui n'a pas encore été interprété ni mis en contexte. Ceci distingue la donnée de l'information, qui est une donnée interprétée, grâce à une mise en contexte. Enfin, lorsque non seulement la donnée est interprétée, mais encore « comprise », c'est-à-dire assimilée et utilisée en vue d'une décision puis d'une action, l'information devient savoir ou connaissance. Les données « tirent leur force » et leur valeur des opérations de production, circulation, agrégation et mise en forme dont elles font l'objet. La transformation des données en information et connaissance nécessite un traitement que réalisent des algorithmes.

Un algorithme est une suite finie d'opérations ou d'instructions, déterminée de manière non ambiguë et permettant de résoudre un problème ou d'obtenir un résultat. La notion de « problème » s'entend ici dans le sens très large d'une tâche à effectuer, comme trier des objets, assigner des ressources, transmettre des informations, traduire un texte, etc. L'algorithme reçoit des données en entrée, par exemple la liste des objets à trier, la description des ressources à assigner, celle des besoins à satisfaire, les informations à transmettre et l'adresse du destinataire, ou encore le texte à traduire, et il produit en sortie d'autres données, par exemple la liste des objets triés, les associations ressources-besoins, un compte-rendu de transmission, la traduction du texte soumis en entrée, etc.

Les algorithmes, programmés sous la forme de logiciels informatiques, sont omniprésents dans le monde numérique. Ils sont les compléments nécessaires des données, leur fonction première étant de trier, d'organiser, de traiter et de présenter les données, de transformer celles-ci de matière première brute en produit élaboré, afin qu'elles puissent répondre aux besoins des utilisateurs. Premières manifestations de l'intelligence artificielle, dont la généralisation constitue aujourd'hui un temps fort de la transition numérique, les algorithmes sont des agents intelligents, des robots, dont il convient évidemment de s'assurer qu'ils opèrent au service d'un intérêt général convenablement et clairement défini, et non pas d'intérêts particuliers et opportunistes.

La mise en données du monde

Au cours des dernières années, la société et l'économie numériques ont connu de profonds bouleversements. Ceux-ci ont affecté les modalités de collecte et d'utilisation des données à caractère personnel, avec des transformations qui sont à la fois techniques, économiques, sociales et culturelles.

D'un point de vue technique, on assiste à la mise en données de nos sociétés (Van Dijck, *voir p.20)*, à la fois des personnes - leurs informations personnelles et identifiantes, leurs préférences et goûts, leur santé et bien-être - et des objets : leurs fonctions et caractéristiques et leur capacité à se connecter à d'autres objets. Ce phénomène est rendu possible par la confluence de plusieurs dynamiques : la progression exponentielle des moyens techniques de captation, de stockage, de reproduction et d'analyse des données ; l'explosion du volume des données qui transitent par ces infrastructures

(mégadonnées ou big data) ; l'essor de l'internet des objets et de l'intelligence artificielle. Au plan économique, de nombreuses entreprises ont vu leurs stratégies de croissance et modèles d'affaires profondément transformés par une logique de valorisation intensive des données personnelles disponibles, permettant : d'un côté, un ciblage publicitaire fin des consommateurs, sous la forme de publicités contextuelles, personnalisées ou comportementales ; et, de l'autre, la production de services personnalisés, souvent plus performants ou novateurs.

Nombre de dynamiques sociales et culturelles se rattachent à ce mouvement, notamment l'adoption rapide des usages fixes et mobiles d'internet sur différents supports, la multiplication des pratiques de partage d'informations, d'expressions, d'opinions, ainsi que la publication de documents personnels sur des plateformes dédiées ou sur les réseaux sociaux. Ces transformations ont redéfini le périmètre des données à caractère personnel et modifié les attitudes des différents acteurs vis-à-vis de la protection de ces données.

L'ère numérique est caractérisée par une omniprésence des données personnelles dans toutes les activités humaines, qu'elles soient publiques, professionnelles ou privées. L'individu se situe au centre d'un réseau complexe d'informations personnelles, relayées par des objets communiquant de plus en plus entre eux de façon autonome, comme le téléphone portable, les bracelets électroniques et plusieurs types de dispositifs électroniques, notamment de vidéosurveillance. La mise en données du monde s'étend ainsi à notre environnement, ce qui tend à brouiller les frontières entre l'univers physique et l'univers en ligne et change le rapport entre données personnelles et vie privée. La protection des unes et de l'autre se superposaient auparavant ; mais, sous l'effet des nouveaux comportements et usages, des nouvelles technologies et modèles d'affaires, la séparation entre la vie privée et la vie publique est aujourd'hui moins clairement identifiée. Une zone grise se dessine, dans laquelle les individus se servent de leurs données personnelles pour mener une vie publique, renoncent à l'idée de la vie privée vue comme un isolement et, tout en demandant une protection, qualifient plutôt celle-ci de maîtrise sur les outils de partage. (Susan Barnes, *voir p.20*) a appelé « paradoxe de la vie privée » le phénomène selon lequel, dans l'environnement numérique, propice à la mise en scène, à l'exposition et à la publicisation de soi-même, les individus, dans la conviction que les technologies numériques participent à la construction de leur personnalité et à leur valorisation sociale et professionnelle, exposent leur vie privée en échange de services ou d'avantages, sans toutefois renoncer à un haut niveau de protection et de maîtrise sur leurs données personnelles.

Le secteur audiovisuel n'échappe pas à ces dynamiques, car le numérique propose un accès multiple et immédiat à une variété de contenus sur une variété de supports, et de plus en plus, les modalités de cette offre foisonnante sont mises en relation étroite avec les préférences et les goûts que l'utilisateur a pu montrer dans son historique d'usage. La personnalisation de l'offre dans le secteur audiovisuel ne se déploie pas sans que les phénomènes de paradoxe de la vie privée et de valorisation de ses données personnelles ne se développent en parallèle.

LES DONNÉES MASSIVES : PUISSANCE, VALEUR ET RISQUE

Les 3 V : Volume, Variété, Vélocité

La législation française sur l'encadrement des traitements de données à caractère personnel répond à la nécessité de réguler le recueil de données aux fins de constitution de fichiers sur les individus. Désormais, un nombre croissant d'acteurs dans le domaine des services internet, notamment les opérateurs de télécommunications les plateformes, les moteurs de recherche et les réseaux sociaux, ont accès à une multitude de données à caractère personnel. Il faut donc prendre en considération les données mises en ligne par les individus eux-mêmes, sur les réseaux sociaux ou les sites de partage, portant sur leur propre vie ou sur celle d'autres personnes, ainsi que **les** traces laissées sur internet, c'est-à-dire les données recueillies automatiquement sous la forme de *cookies* ou de signaux de localisation lors de navigations sur internet. Grâce au recours à la géolocalisation, en particulier, une économie cachée de la collecte des traces permet de suivre, d'analyser, de mesurer et de monétiser l'activité des utilisateurs d'appareils mobiles, pour des objectifs commerciaux et publicitaires. Ces données sont à la fois de formats hétérogènes (images, textes, vidéos…) et d'origines diverses, allant de caractéristiques sociodémographiques « objectives » de la personne à des informations plus subjectives, comme ses goûts, relations ou opinions.

L'expression *big data* ou « mégadonnées » désigne cette collecte massive de données et l'ensemble des instruments algorithmiques qui sont utilisés pour en tirer du sens, c'est-à-dire transformer les données en information puis connaissance. Le *big data* est souvent associé au triptyque des « trois V » : volume, variété et vélocité. Il s'agit respectivement de la masse des données collectées, de leur grande diversité et de la rapidité de leur traitement permise par les nouvelles générations de matériels et d'algorithmes. Il en résulte une démultiplication des possibilités d'analyse et de calcul dans les domaines les plus variés, allant du ciblage publicitaire à la veille sanitaire, du diagnostic médical à la prévention du déclenchement de phénomènes naturels, ou encore, au développement des villes intelligentes (*smart cities*).

La collecte massive de données personnelles est amplifiée par l'organisation industrielle de l'économie numérique, marquée par un foisonnement des rachats et fusions de sociétés. On se rappelle notamment le rachat récent par Facebook des sites de partage de photographies Instagram et de messagerie instantanée WhatsApp. Par ailleurs, on constate que les courtiers en données personnelles (*data brokers*), spécialisés dans la revente des données collectées, prolifèrent.

La puissance inédite de calcul et d'analyse permise par le *big data* converge avec la considérable capacité de stockage offerte par l'informatique en nuage (*cloud computing*), ainsi que l'extrême diversité des données recueillies ou partagées. Cette convergence donne naissance à des systèmes de détection, de classification et d'évaluation anticipative des comportements humains, de « gouvernementalité algorithmique » (Rouvroy et Berns, *voir p.20*). Les prises de décision automatique ou semi-automatique sont de plus en plus fréquentes et visent à anticiper les comportements, les goûts et les choix de chacun. Des techniques telles que l'exploration des données (*data mining*) et le profilage permettent de personnaliser les offres de services sans que les individus concernés n'aient préalablement manifesté des intentions ou explicité des préférences.

La valeur de la donnée dépend de nombreux facteurs

Dans l'économie numérique les opérations de mise en forme des données, sous toutes leurs facettes, ont pour objectifs leur valorisation et leur monétisation. En effet, les entreprises les reconnaissent de plus en plus comme un « *nouvel actif intangible, susceptible de rapporter un revenu direct ou indirect* » (Gomery, *voir p.20*), et voient donc leur monétisation comme une façon de développer de nouveaux services marchands et de transformer leurs modèles d'affaires. Cependant, apparaissent notamment trois questions : comment créer de la valeur en respectant la confidentialité et les engagements pris avec le consommateur ? Comment garder une transparence à la fois sur l'usage qui est fait des données et éventuellement sur les façons de partager équitablement la valeur qui en serait extraite ? Comment créer des synergies et des partenariats au sein de ce marché des données sans impacter négativement la neutralité des plateformes, la concurrence loyale et la propriété intellectuelle ?

Pour les entreprises, il s'agit souvent d'analyser le plus précisément possible leurs audiences en ligne afin de les valoriser auprès des annonceurs. C'est notoirement l'une des principales sources de revenus du réseau social Facebook.

La valeur des données collectées par les entreprises dépend de plusieurs facteurs, notamment le degré de fraicheur et le degré d'exclusivité de la donnée : la fraicheur puisqu'une donnée déjà utilisée peut perdre de son intérêt pour un utilisateur tiers ; l'exclusivité puisqu'une donnée non partagée peut présenter un caractère stratégique pour son détenteur.

Du côté des consommateurs, il est intéressant de noter que la valeur de la donnée n'est souvent pas la même pour eux que pour l'entreprise, et que cette valeur est variable selon les utilisateurs et selon le contexte sociétal, culturel ou géographique.

La valeur de la donnée dépend de la nature de ses consommateurs-utilisateurs mais celle-ci peut varier. De nombreux travaux en psychologie, sociologie et économie comportementale se sont attachés à montrer que les individus accordent une valeur variable à leur protection de données de manière peu ou pas rationnelle. Une étude menée aux États-Unis et à Singapour, révèle que, pour les internautes interrogés, la protection des données personnelles contre les erreurs, les accès non autorisés ou les usages détournés serait valorisée entre 30 et 45 dollars ; cependant, une différence de 2 dollars sur une carte d'achat peut sembler suffisante à une majorité d'entre eux pour compenser une perte d'anonymat dans des transactions en ligne (Acquisti et al., *voir p.20*). Une autre étude (Brandimarte et al., *voir p.20*) démontre que des consommateurs bénéficiant de mécanismes de contrôle de leurs données sur une plateforme numérique sont plus enclins à communiquer des données personnelles sensibles, parce que leur capacité de contrôle les porte à abaisser leur niveau de vigilance. On constate que la plupart des internautes ne lit pas les conditions générales d'utilisation, de format souvent encyclopédique, que pourtant ils acceptent : ainsi, lors d'une expérimentation, 7 500 personnes ont accepté à leur insu de céder leur âme pour l'éternité à un fournisseur de services (Borgesius, *voir p.20*) ! Selon d'autres travaux, le droit au respect de la vie privée se serait néanmoins transformé : initialement droit individuel à être laissé en paix et protégé des

intrusions d'autrui, il tend aujourd'hui à faire l'objet d'une « négociation collective » dont l'objectif principal est de maîtriser la projection de soi dans les interactions sociales avec autrui (Casilli, *voir p.20*).

Les risques pour le consommateur et le citoyen

Les acteurs du numérique, tout particulièrement les grandes plateformes, recourent de plus en plus à des algorithmes prédictifs, destinés notamment à deux objectifs : personnaliser le service rendu aux clients, fournir des éléments utiles à une prise de décision. En raison de leur nature technique, on a tendance à attribuer aux algorithmes des qualités d'infaillibilité et d'objectivité, alors même qu'ils portent l'empreinte des partis initialement pris par leurs développeurs. Il s'agit en effet de procédures codées pour transformer des données « entrantes » en données « sortantes », sur la base de calculs spécifiques. Ces dispositifs peuvent en partie déposséder les individus des choix qu'ils pourraient faire spontanément et ainsi réduire leur libre arbitre, en n'agissant pas simplement comme des mécanismes d'aide à la décision mais comme de véritables systèmes de décision automatique ou semi-automatique. Les algorithmes peuvent également altérer les droits des consommateurs et modifier la relation de l'usager-citoyen avec les pouvoirs publics. Parallèlement au déploiement à large échelle des algorithmes par le secteur privé, émerge une véritable « action publique algorithmique », par exemple destinée, à la manière de ce que montre le film *Minority Report*, à anticiper des comportements criminels en vue de renforcer la sécurité ; d'autres applications se rapportent au domaine social, telles que la prédiction des risques de maltraitance, l'aide au diagnostic médical ou l'anticipation des risques de décrochage scolaire…

Se pose dès lors la question d'intégrer les droits d'information, d'accès et d'opposition, déjà reconnus aujourd'hui, au sein d'un encadrement accru des algorithmes employés pour les traitements de données personnelles. L'article 10 de la loi Informatique et libertés interdit d'ores et déjà qu'une décision portant des effets juridiques à l'endroit d'un individu soit prise « *sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité* ». Toutefois, à l'ère de la « gouvernementalité algorithmique » (*cf. supra*), cette interdiction semble de moins en moins opérante *de facto*, au vu de l'importante asymétrie de situation entre les personnes soumises à l'action algorithmique et les concepteurs des algorithmes, ces derniers détenant un double avantage en matière d'information et de décision.

De nouvelles règles sont par conséquent envisagées, aux niveaux national et international, pour encadrer le fonctionnement des algorithmes prédictifs. Plusieurs sont concevables : un droit d'opposition au profilage pouvant conduire à des mesures produisant des effets juridiques pour l'individu, ou affectant de manière importante ses libertés, intérêts et droits ; une contrainte d'intervention humaine effective sur toute décision prise à l'aide d'un algorithme ; une obligation de transparence sur les types de données personnelles utilisées par l'algorithme et les modalités de son paramétrage, conduisant, le cas échéant, à des audits qui permettraient de contester la logique générale de l'outil ou la véracité des données analysées ; l'interdiction d'utiliser des algorithmes conduisant, directement ou indirectement, à instaurer une

discrimination fondée, notamment, sur la race, l'origine ethnique, les opinions politiques, la religion, les convictions, l'orientation sexuelle ou l'identité de genre, ou encore l'appartenance syndicale.

Au-delà de son impact immédiat sur la vie privée, la généralisation de la prédiction et de la personnalisation algorithmiques affecte à moyen terme l'organisation sociale dans son ensemble, soulevant ainsi une large variété de questions, dont le risque d'une discrimination accrue, d'une uniformisation de la société, d'une réduction de la liberté des choix individuels, voire d'une manipulation, lorsque la personnalisation a pour but d'influencer les comportements.

Le secteur de l'audiovisuel plongé dans l'univers des algorithmes

Différentes familles d'algorithmes

Dans le secteur audiovisuel, les algorithmes sont nombreux et variés : moteurs de recommandation, publicité programmatique, programmation des radios, construction de fils conducteurs cross-média, en constituent autant d'exemples. On peut, en première approche et de manière non exhaustive, distinguer les algorithmes de mise en correspondance (*matching*), les algorithmes d'éditorialisation, les algorithmes de profilage ou de ciblage. Dans la suite de cette section, l'attention sera portée en particulier sur les algorithmes de *matching* que sont les moteurs de recommandation-prescription.

Les moteurs présents sur les SMAD[1], les distributeurs OTT[2] et les plateformes numériques (sites de partage ou réseaux sociaux), sont basés sur trois principes algorithmiques distincts, pouvant éventuellement être combinés entre eux.

- *Algorithme-menu* : à ce degré minimaliste de l'algorithmique, les contenus proposés sont classés *ex ante* par genres et sous-genres, selon une typologie prédéfinie par l'opérateur, et l'utilisateur choisit *ex post* au sein de ce menu. Le distributeur OTT Molotov correspond typiquement à ce modèle.
- *Algorithme statistique* : les contenus sont « poussés » vers l'utilisateur, en fonction de ses propres consommations passées, ainsi que des contenus aimés par ses amis sur les réseaux sociaux. Netflix, par exemple, recourt à un algorithme de ce type.
- *Algorithme sémantique* : la base de contenus constituant le catalogue est analysée en profondeur et indexée à l'aide de mots-clés. L'utilisateur entre lui-même, de son côté, des mots-clés caractérisant ses goûts. L'algorithme établit alors une correspondance entre les mots-clés descriptifs de l'offre et ceux descriptifs de la demande. L'outil développé par la société Spideo[3] est notamment fondé sur ce principe.

[1] Les services de médias audiovisuels à la demande ou SMAD, comme la vidéo à la demande à l'acte ou par abonnement, ou encore la télévision de rattrapage, permettent de visionner des programmes audiovisuels au moment choisi par l'utilisateur, sur le poste de télévision ou sur l'ordinateur, de manière gratuite ou payante. Source : www.csa.fr.

[2] Un service over-the top, ou OTT, est un service de livraison d'audio, de vidéo et d'autres médias sur internet sans la participation d'un opérateur de réseau traditionnel (fournisseur d'accès à internet, câblo-opérateur ou opérateur satellitaire) dans le contrôle et la distribution du contenu. Source : Wikipédia.

[3] Spideo est une startup spécialiste de la recommandation de contenus vidéo. À partir d'une approche sémantique de la recommandation, Spideo a développé des logiciels de découverte de films, séries et programmes de télévision, commercialisés auprès des acteurs de la vidéo à la demande et de la télévision interactive.

L'algorithme-menu peut en réalité être considéré comme une version élémentaire de l'algorithme sémantique, où l'indexation des contenus est réduite à un classement typologique et où l'expression des goûts de l'utilisateur se borne à une navigation au sein de ce classement. Il n'existe par conséquent, en définitive, que deux logiques directrices contrastées, respectivement statistique et sémantique.

Tenir compte des affinités tout en préservant la découverte

La structure des goûts d'un consommateur peut s'analyser comme un réseau abstrait. Ce réseau comporte *des liens forts*, traduisant une préférence avérée pour un type de contenus bien identifié *a priori* ; *des liens faibles*, rendant compte d'une affinité non encore révélée pour un type de contenus restant à découvrir *a posteriori*.

Un algorithme bien tempéré devrait respecter la structure du graphe des affinités, c'est-à-dire réaliser un juste équilibre entre exploitation des acquis indiqués par les liens forts et exploration des potentialités nouvelles indiquées par les liens faibles. La recherche de ce juste équilibre correspond à un optimum de performance stratégique du consommateur dans la satisfaction dynamique de ses préférences, selon un résultat connu de théorie des jeux se rapportant à l'analyse des jeux répétés dans le temps

Par construction même, les algorithmes statistiques n'utilisent que les liens forts, ceux qui sont créés et activés par les consommations passées et courantes, et ils ont en outre pour effet d'encore renforcer ces liens, sacrifiant ainsi l'exploration au profit de l'exploitation. Les algorithmes sémantiques laissent au contraire une chance à l'exploration, en proposant des contenus qui, certes, sont conformes aux goûts déclarés par l'utilisateur mais qui ne sont pas nécessairement semblables à ses consommations passées.

Néanmoins, ni les algorithmes statistiques, ni les algorithmes sémantiques ne prennent en compte « l'incertitude créatrice » et la « sérendipité » (ou flânerie en ligne), négligeant ainsi la dimension aléatoire des choix de consommation. Si l'algorithmique statistique me conduit à aimer ce que j'aime déjà et l'algorithmique sémantique, à aimer ce que je pourrais aimer, comment m'amener à aimer ce que j'ignore encore pouvoir aimer ? À cet effet, il faudrait imaginer un algorithme statistique qui, avec une probabilité faible mais strictement positive, inverserait le sens de son critère de distance pour m'emmener au plus loin, et non pas au plus près, de mes choix antérieurs ; ou bien, imaginer un algorithme sémantique soumis à de petites perturbations qui, parfois, ajouterait ou supprimerait certains éléments de mon profil déclaré, afin de m'orienter, grâce à cet « aléa contrôlé », vers de nouveaux territoires cognitifs.

La prise en main des outils par l'utilisateur

Le phénomène *big data* et la montée en puissance des algorithmes sont clairement révélés dans le langage courant par l'usage grandissant du qualificatif *smart*, employé pour désigner les biens et les services issus de l'économie numérique et, plus généralement, tout « objet » prénumérique enrichi de fonctionnalités dites intelligentes, allant du *smartphone* et de la *smart TV* à l'échelle micro, jusqu'à la *smart city* à l'échelle macro.

Les algorithmes, qui renferment l'intelligence numérique, constituent, en symbiose avec les données qu'ils consomment et celles qu'ils produisent, les premiers artisans du *smart*. Le monde du *smart* est ainsi celui que façonnent pour nous les algorithmes et leur halo de données entrantes et sortantes. Mais le bien-être que devrait nous promettre un « développement numérique durable » réside-t-il véritablement dans le *smart* ? Ne réside-t-il pas plutôt dans le

cool ? Le *smart* c'est quand on m'impose selon une logique *top-down* un service prédéfini et censé me satisfaire au mieux, le *cool* c'est quand je compose au moins en partie ce service moi-même, selon une logique *bottom-up*.

L'arbitrage entre le *smart* et le *cool* renvoie aux débats sur la doctrine du *soft computing* : jusqu'à quel degré la robotique doit-elle se faire oublier, permettant ainsi une fluidité accrue de l'usage au prix d'une certaine opacité, ou doit-elle inversement dévoiler ses rouages, permettant ainsi un plus grand contrôle par l'utilisateur, au prix d'un coût de participation plus élevé ? Dans la confrontation du *cool* et du *smart*, l'exercice d'un plus grand libre arbitre se mérite dans l'effort collaboratif ! Le *smart* satisfait mieux les consommateurs passifs et dociles, du moins tant qu'ils ne se sentent pas manipulés, tandis que le *cool* plaît davantage aux consommateurs actifs, ou « consommacteurs ».

L'algorithmique statistique est l'archétype du *smart* : elle propose au consommateur ce qui est supposé lui plaire le plus, sans lui laisser la main dans le processus de sélection. L'algorithmique sémantique relève quant à elle davantage du registre du *cool* (tel qu'il a été défini plus haut), invitant le consommateur à participer au choix de la proposition, sur le mode d'une interaction avec un système expert proche de l'humain (*human like)*.

Les données et les algorithmes, instrument d'un développement numérique durable

Maintenir et, si possible, accroître la diversité culturelle et la santé du secteur audiovisuel, assurer la protection des consommateurs, développer la confiance en la société numérique : comment les algorithmes et l'utilisation de données se comportent-ils au regard de ces trois grands objectifs d'un développement numérique durable ?

Préserver la diversité culturelle

L'utilisation des algorithmes par les industries culturelles contribue à la diversité en facilitant la découverte d'œuvres audiovisuelles qui ne seraient pas par ailleurs programmées en raison de leur petit budget, ou en raison de l'absence d'un distributeur ou d'un budget de promotion. Ainsi, grâce aux moteurs de recommandation, certains films peuvent trouver un public même si ces films ne sont pas programmés par les chaînes de télévision traditionnelles.

Mais ces algorithmes peuvent également conduire à des effets inverses, à savoir enfermer les individus dans une personnalisation des services en fonction de leurs goûts et opinions. Si tel était le cas, il en résulterait potentiellement une atteinte au libre choix, une homogénéisation de l'information, une polarisation des contenus autour de visions dominantes en opposition avec l'objectif de diversité culturelle. Le recours massif et automatique aux algorithmes nécessite donc une vigilance, pour assurer que les côtés potentiellement nocifs ne prennent pas le dessus sur les côtés positifs.

Parce qu'ils n'excluent pas totalement la prise en compte des liens faibles, les algorithmes sémantiques sont vraisemblablement les plus favorables à la diversité culturelle mais ils sont aussi les plus coûteux à mettre en œuvre, exigeant en effet un lourd travail préalable d'indexation de la base de contenus. Les algorithmes statistiques, à travers leur dynamique d'autorenforcement des liens forts, créent à l'inverse un risque d'enfermement du consommateur dans une bulle culturelle figée. En contrepartie, ces algorithmes

sont relativement peu coûteux à implémenter puisque les données qu'ils utilisent, déjà disponibles en ligne, n'ont pas à être construites : elles ne sont autres que les traces numériques laissées par les internautes-consommateurs lors de leurs parcours sur les sites de fourniture de contenus et de leur fréquentation des réseaux sociaux.

L'utilisation des algorithmes par les industries créatives peut avoir des effets plus ou moins importants sur le développement d'une offre culturelle riche et diversifiée. Ces effets – positifs ou négatifs – ne sont pas tant le fait de leur mobilisation massive par les fournisseurs de SMAD, plateformes numériques et distributeurs OTT, que de leur nature et de la manière dont ils sont perçus par les consommateurs. Aussi, le risque porté sur la diversité culturelle réside moins dans l'hégémonie des algorithmes que dans la domination d'un acteur monopolistique dont l'algorithme conduirait aux effets d'enfermement évoqués ci-dessus et dans la difficulté pour l'utilisateur de reconstruire *a posteriori* le processus de la recommandation algorithmique.

Afin de préserver le libre choix de l'individu et de lui permettre de contextualiser les recommandations personnalisées de contenus qui lui sont faites, il est nécessaire d'assurer un certain degré d'information du consommateur sur les mécanismes directeurs des algorithmes, notamment les critères de sélection mis en œuvre pour produire les résultats, en distinguant les suggestions « naturelles », celles qui sont déduites sans biais de l'algorithme, des suggestions commercialement sponsorisées. Cet objectif pourrait être en partie atteint par la publication d'indices mesurant le caractère ouvert ou fermé d'un algorithme vis-à-vis de différents critères de recherche, publication qui inciterait le marché à adopter des pratiques vertueuses. Par ailleurs, une obligation stricte de transparence pourrait être imposée si l'algorithme tenait compte d'un paiement ou autre avantage accordé par un partenaire commercial quelconque. Ces dispositions, visant à entrouvrir la « boîte noire algorithmique » sans violer le secret commercial ni décourager l'innovation, apparaissent comme une contrepartie naturelle à la captation et à la modélisation permanente des comportements des utilisateurs.

Les algorithmes sémantiques, parce que l'utilisateur y introduit lui-même ses propres données de profil, apparaissent comme relativement « transparents », par comparaison aux algorithmes statistiques qui élaborent des profils de manière le plus souvent non explicites, à partir de comportements observés largement à l'insu de l'utilisateur, même dans le cas où le consentement préalable de celui-ci a été « formellement » recueilli. Par ailleurs, du point de vue de la maîtrise des données personnelles (*privacy*), la logique sémantique domine également la logique statistique, en ce qu'elle exige une participation active de la part de l'utilisateur, transférant vers celui-ci une part de responsabilité dans la construction de son profil. De plus, ce profil est à tout moment modulable et révisable

Promouvoir une maîtrise responsable des données

L'industrie audiovisuelle utilise des données pour cibler la publicité depuis bientôt un siècle. Introduit dans les années 1920, le système Nielsen Ratings[1] et son équivalent Médiamétrie en France, sont des outils d'analyse de données personnelles qui permettent de créer un marché efficace pour la publicité. Sans ces outils, le marché publicitaire pour le secteur audiovisuel

[1] Wikipedia, https://en.wikipedia.org/wiki/Nielsen_ratings.

serait beaucoup moins performant, et le financement de la création en souffrirait.

Cependant, l'évolution dans la captation et l'utilisation massive des données comportementales à des fins de publicité et de personnalisation du service pose des questions relatives à la protection des droits fondamentaux. C'est un enjeu d'autant plus important que les données de consommation de contenus culturels comportent une fonction identitaire très forte. La consommation de biens et de services culturels touche à notre identité subjective, mais aussi à notre identité sociale et communautaire, c'est à dire à l'image que nous souhaitons renvoyer dans notre environnement social. Par ailleurs, les individus sont moins susceptibles de mettre en place des stratégies de dissimulation dans leur manière de consommer les contenus culturels que sur les réseaux sociaux. Les utilisateurs ont en effet davantage d'incitations à mettre en avant ce qu'ils valorisent ou à cacher ce qui les dérange sur Facebook ou Instagram. De ce fait, l'exploitation des données relatives aux contenus culturels est particulièrement sensible. Elle doit alors être encadrée, selon des modalités à la fois protectives et proactives. Il s'agit en effet non seulement de nous protéger d'irruptions non souhaitées dans nos sphères d'intimité mais également de nous permettre de connaître les biais qui peuvent être introduits par la collecte d'informations nous concernant, et de choisir en conséquence les éléments que nous souhaitons partager.

Ce droit à une autodétermination informationnelle répond à des aspirations tant personnelles que collectives. Il permet en effet à l'individu d'agir sur l'image qu'il renvoie et sur sa réputation sociale. Dans une logique cumulative, il permet également de redonner à l'ensemble des utilisateurs de services un pouvoir d'agir plus fortement sur la nature de l'offre culturelle qu'ils reçoivent. Cela en leur permettant de moduler les informations qui peuvent être exploitées par les acteurs économiques pour inférer sur l'état de la demande. Par ailleurs, si l'on considère que la consommation de contenus culturels en ligne peut s'apparenter à un travail implicite (*digital labor*) derrière un objectif de loisir, la reconnaissance du droit des individus de maîtriser ces données participe à une reconnaissance par la société de cette activité productive des consommateurs.

Cette capacité à maîtriser ses données se traduit par la nécessité pour l'individu d'accéder à des informations intelligibles sur les utilisations qui en sont faites. Le nouveau règlement européen relatif à la protection des données à caractère personnel (RGPD) réaffirme en ce sens le droit d'accéder aux données qui sont collectées sur soi et le droit d'être informé sur les traitements qui seront faits à partir d'elles. L'activation effective de ces droits passe notamment par la mise à disposition de conditions générales d'utilisation plus lisibles : moins longues, moins ambiguës et compréhensibles par des personnes non expertes. D'autres initiatives permettant de faciliter la gestion de leurs données par les individus gagneraient à être explorées, telles que la fourniture de tableaux de bord pour le suivi et la sélection de l'utilisation des données et le paramétrage de la personnalisation du service.

Le droit à la portabilité des données, introduit par le RGPD et par la loi pour une République numérique, s'inscrit dans la philosophie de l'autodétermination informationnelle, en nous permettant de récupérer nos données sous un format standardisé et de les réutiliser. L'objectif est notamment de permettre à l'utilisateur de ne pas être enfermé dans un écosystème captif et de faire lui-même usage de ses données à des fins personnelles ou bien pour les partager vers d'autres services. La portabilité des données constitue à ce titre un moyen pour lutter

plus efficacement contre les restrictions de concurrence au détriment de l'individu et des autres acteurs du marché.

Cependant ce droit ne s'étend pas jusqu'à la possibilité d'accéder aux raisonnements et informations obtenues à partir du traitement de nos données brutes par des algorithmes (cf. définitions des données et informations proposées en introduction). La traçabilité de l'origine de toutes ces informations peut en effet s'avérer complexe lorsque les traitements des algorithmes sont appliqués sur des bases de données volumineuses et/ou devenues anonymes. Nous sommes donc, d'un côté, informés sur les données brutes qui sont collectées à notre sujet (conditions générales d'utilisation et droit à la portabilité) et, d'un autre côté, informés de manière synthétique sur notre profil d'utilisateur. Nous n'avons pas accès au lien de causalité, au raisonnement qui a fait passer de ces données brutes à ce profil. Il est néanmoins envisageable de permettre aux individus, aux collectifs qui défendent leurs intérêts et à la recherche d'inférer à leur tour sur les raisonnements pratiqués par les plateformes pour catégoriser leurs utilisateurs et leur recommander des contenus. Cela en combinant des droits individuels (information, accès, rectification, portabilité) à des démarches collectives : recherches, mise en commun de données, nouveaux services, etc.

Garantir la loyauté des algorithmes

Les données issues de la consommation de contenus culturels et ludiques peuvent s'avérer particulièrement sensibles en ce qu'elles révèlent par elles-mêmes sur la personne, sur ses affinités, ses intérêts, mais aussi en ce qu'elles peuvent permettre d'inférer sur son mode de vie. Pour ces raisons et afin de préserver la confiance entre les utilisateurs et les fournisseurs de services, il est essentiel que ces données ne fassent pas l'objet de détournements en usages marketing ou publicitaires cachés ou discriminatoires – volontairement ou non –, mais qu'elles servent bien à la personnalisation du service dans l'intérêt de l'ensemble des parties. Une utilisation des données pour la publicité reste importante afin de permettre au secteur audiovisuel de continuer à assurer son financement, mais cette utilisation accrue de données doit s'accompagner de protections pour l'individu.

Le principe de loyauté vise à obliger les acteurs économiques à assurer de bonne foi les services qu'ils proposent sans chercher à les détourner à des fins contradictoires à l'intérêt de leurs utilisateurs. En ce qui concerne les algorithmes de personnalisation, ce principe vise principalement à respecter un principe général de non-discrimination et à rendre plus transparents les modes de collecte, de traitement des données et de restitution de l'information. La loyauté des algorithmes participe de fait à créer davantage de confiance, facteur essentiel de l'acceptabilité sociale des technologies numériques. Cette obligation de loyauté a été introduite par la loi pour une République numérique et s'applique aux plateformes qui mettent à disposition des contenus audiovisuels fournis par des tiers.

L'algorithmique statistique, qui repose sur la « filature » électronique comme moyen de profilage, n'inspire guère la confiance, au contraire de l'algorithmique sémantique, qui crée les conditions d'un dialogue homme-machine simulant un dialogue réel entre un utilisateur et un expert. Tandis que l'algorithmique statistique ne fait que recycler une information purement endogène à la sphère de l'utilisateur, l'algorithmique sémantique apporte une expertise exogène, incorporée dans l'indexation préalable des contenus ; une expertise « loyale », assez similaire aux conseils « loyaux » que pouvait prodiguer à l'ère prénumérique un exploitant cinéphile de vidéo shop. Les deux types d'algorithmes peuvent utilement coexister.

Le développement des algorithmes est un domaine où l'innovation avance très vite. Toute tentative de réguler les algorithmes serait probablement vouée à l'échec en raison de l'évolutivité de la technologie et du caractère confidentiel et concurrentiel des développements. Cependant, le régulateur pourrait éventuellement contribuer à la transparence du marché en publiant des indices mesurant différents aspects qualitatifs des algorithmes. Ces indices pourraient éventuellement mesurer le niveau de contribution des algorithmes aux différents objectifs de la politique audiovisuelle, tels que la diversité.

Annexe

Algorithme : quelle étymologie ?

C'est le savant perse Al-Khwarizmi, dont le nom a été latinisé en Algoritmi, qui est à l'origine du mot algorithme. Né vers 780 dans la région du Khwarezm située dans l'actuel Ouzbékistan et d'où il tire son patronyme, mort vers 850 à Bagdad, Al-Kwharizmi a entièrement vécu sous la dynastie abbasside. Membre de la « Maison de la sagesse » de Bagdad, il était tout à la fois mathématicien, géographe, astrologue et astronome.

Ses travaux, rédigés en langue arabe, puis traduits en latin à partir du 12e siècle, sont à l'origine de l'introduction des chiffres arabes et de l'algèbre en Europe. Son ouvrage le plus célèbre, intitulé Kitābu al-mukhtaṣar fī ḥisābi al-jabr wa'l-muqābalah (كتاب المختصر في حساب الجبر والمقابلة), soit *Abrégé du calcul par la restauration et la comparaison*, est aujourd'hui considéré comme le premier manuel d'algèbre.

Al-Khawarizmi a classifié les algorithmes existants à son époque, notamment selon leurs critères de terminaison, mais il n'est pas l'inventeur de ces outils : l'algorithme sans doute le plus célèbre, celui d'Euclide pour le calcul du PGCD, est très antérieur. Par ailleurs, l'usage pratique de divers algorithmes s'était précédemment développé en Mésopotamie, notamment pour réaliser les calculs complexes de l'impôt à Babylone !

Si le mot « algorithme » s'écrivait « algorythme», alors son étymologie virtuelle, toute différente, ne devrait plus rien au Perse Al-Khwarizmi. Du grec ancien *algos*, douleur, et de *rhuthmos*, mouvement balancé, il résulterait que *algo-rhuthmos* signifierait « cadence douloureuse » ou « douleur lancinante ». Sur une pareille promesse, qui voudrait d'un environnement numérique peuplé « d'algorythmes » ? Aux « algorythmes », on préférerait très certainement les « euphorythmes », troquant la douleur (*algos*) d'une passivité opaque et subie, contre le plaisir de se bien porter (*euphoros*), grâce à une participation active et éclairée de l'individu dans une co-activité homme-machine.

Références

1 José Van Dijck (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197.

2 Susan B. Barnes (2006). *A privacy paradox: Social networking in the United States*. First Monday, 11(9).

3 Antoinette Rouvroy et Thomas Berns, « Le nouveau pouvoir statistique », *Multitudes* n° 40, pp. 88-103, 2010.

4 Reda Gomery (2015). « Monétisation des données : enjeux et limites d'une pratique en émergence », Blog D.Views, http://www.blog.deloitte.fr/economie/monetisation-des-donnees-enjeux-et-limites-dune-pratique-en-emergence/.

5 Alessandro Acquisti, Laura Brandimarte, George Loewenstein, « Privacy and human behavior in the age of information », *Science* 30 January 2015, vol. 345, n° 6221, pp. 509-514.

6 Laura Brandimarte, Alessandro Acquisti et George Loewenstein, *Misplaced Confidences: Privacy and the Control Paradox*, 2010.

7 Frederik J. Zuiderveen Borgesius, *Consent to behavioural targeting in european law - What are the policy implications of insights from behavioural economics ?*, 2013.

8 Antonio Casilli, *Quatre thèses sur la surveillance numérique de masse et la négociation de la vie privée*, Rapport du Conseil d'État, 2015, pp. 423-434.

9 Latanya Sweeney, *Weaving technology and policy together to maintain confidentiality*, *Journal of Law, Medicine and Ethics*, 25, 1997, pp. 98–110.