Le deep learning au service de la prédiction de l'orientation sexuelle dans l'espace public

Nicolas Baya-Laffite, Boris Beaude, Jérémie Garrigues

LE *DEEP LEARNING* AU SERVICE DE LA PRÉDICTION DE L'ORIENTATION SEXUELLE DANS L'ESPACE PUBLIC

Déconstruction d'une alerte ambiguë

Nicolas BAYA-LAFFITE
Boris BEAUDE
Jérémie GARRIGUES

DOI: 10.3917/res.211.0137

En septembre 2017, plusieurs médias se font l'écho d'une alerte lancée par deux chercheurs de la Graduate School of Business de l'Université de Stanford, le professeur en psychologie et *data science* Michal Kosinski et le chercheur en *machine learning* Yilun Wang. Dans le pre-print de la recherche qui engage cette alerte[1], Kosinski et Wang font état de nouveaux risques pour la vie privée découlant de la mise au point d'un modèle algorithmique qui parviendrait à identifier l'orientation sexuelle à partir d'images de visages. En exploitant les données produites par un réseau de neurones appliqué à la reconnaissance faciale, les deux chercheurs auraient mis au point un classificateur binaire de l'orientation sexuelle très performant, susceptible d'être appliqué à l'espace public. Le titre de l'article alors à paraître dans le *Journal of Personality and Social Psychology* est particulièrement explicite : « Deep neural networks are more accurate than humans at detecting sexual orientation from facial images » (Wang et Kosinski, 2018)[2]. Trois lignes de préoccupation seront à l'origine de l'alerte.

En premier lieu, le protocole expérimental mis en place par Kosinski et Wang a pour but de démontrer que le visage dévoile des informations sur l'orientation sexuelle invisibles à l'œil humain, mais perceptibles par une machine. Selon les auteurs, les progrès du *deep learning* sont tels que le champ des possibles laisse place à une réhabilitation de la physiognomonie, qui avait perduré au tournant du XX^{e} siècle sous la forme de la phrénologie, et persisté tout au long du XX^{e} siècle via la morphopsychologie, et dont on sait qu'elle a toujours été suivie et accompagnée de peurs et d'inquiétudes (Le Breton, 1992 ; Fara, 2008).

---

1. Soumis au *Journal of Personality and Social Psychology* le 4 juin 2017, et après la réception d'une révision le 29 juillet 2017, l'article est accepté pour publication le 12 août 2017 et se trouve disponible depuis février 2018 sur le site de l'*American Psychological Association*. Le pre-print de l'article est accessible sur le volet « files » site web compagnon de l'étude créé en 2016 sur la plateforme *Open Science Framework* (OSF). Toutes les versions (10 au total entre le 2 septembre et le 24 septembre 2017) sont disponibles à l'adresse https://osf.io/fk3xr/?show=revision, consulté le 05/10/2018.
2. Signé Yilun Wang et Michal Kosinski, nous nous référerons pourtant aux auteurs du papier dans l'ordre inverse, Kosinski et Wang. Si les deux auteurs déclarent dans l'article avoir contribué à parts égales à l'article, c'est bien Michal Kosinski le chercheur senior qui en assume la responsabilité en tant que *corresponding author*.

En second lieu, s'inscrivant dans un renouveau des recherches sur le « gaydar », c'est-à-dire les capacités des individus à reconnaître l'orientation sexuelle à partir d'un visage[3] (Gelman *et al*., 2018 ; Miller, 2018), Kosinski et Wang ont recours à la *théorie de l'exposition prénatale aux hormones* (*Prenatal Hormones Theory* ou PHT) afin d'expliquer pourquoi l'orientation sexuelle s'exprimerait dans les traits du visage. Cette théorie fait de l'orientation sexuelle un élément déterminé pendant la gestation et dès lors irréversible. Bien que s'insérant dans un large faisceau de travaux initiés il y a plusieurs décennies sur le rôle des hormones prénatales, la PHT n'en reste pas moins controversée (Bailey *et al*., 2016 ; Jordan-Young, 2011). Qui plus est, avec la PHT, Kosinski et Wang reconduisent une « biologisation » des origines de l'orientation sexuelle, dont la désirabilité s'est beaucoup réduite ces dernières années (Jannini *et al*., 2010). La PHT n'en reste pas moins essentielle à l'alerte lancée par Kosinski et Wang. Les androgènes déterminant l'orientation sexuelle seraient en effet aussi susceptibles d'avoir des effets sur la physionomie voire la morphologie des individus, affectant par exemple la couleur de la peau, la pilosité et plus encore la largeur de la mâchoire. C'est pourquoi les visages porteraient la trace de l'orientation sexuelle. La performance prédictive du modèle reposant ainsi sur des dispositions naturelles, et par continuité transculturelle, la possibilité de généraliser les performances en laboratoire justifierait la pertinence de l'alerte.

Enfin, compte tenu de l'actualité des applications à l'espace public de technologies de reconnaissance faciale dans un contexte de surveillance de masse (Introna et Nissenbaum, 2010 ; Klauser, 2016 ; Norval et Prasopoulou, 2017), la vie privée s'en trouverait affectée et, en particulier, la vie privée des personnes homosexuelles. Estimant que des méthodes similaires de profilage facial à base de réseaux de neurones seraient utilisées par des entreprises et des gouvernements (Lubin, 2016), Kosinski et Wang justifient la publication des résultats au titre d'une alerte qui serait d'autant plus sérieuse à l'heure des médias sociaux et de la multiplication des caméras de surveillance, dans un contexte mondial toujours marqué par l'homophobie[4]. En interpellant les

---

3. Mélange de « gay » et de « radar », ce mot argotique est utilisé, y compris dans la littérature scientifique, pour se référer à la capacité des individus à identifier une personne homosexuelle par intuition ou en interprétant des signaux subtils véhiculés par l'apparence ou le comportement.

4. Dans les termes des auteurs : « We did not create a privacy-invading tool, but rather showed that basic and widely used methods pose serious privacy threats. We hope that our findings will inform the public and policymakers, and inspire them to design technologies and write policies

citoyens, Kosinski et Wang s'inscriraient en cela dans la récente inflation des alertes concernant le numérique, dont Edward Snowden, avec sa dénonciation des écoutes généralisées, constitue sans doute l'exemple le plus connu dans le domaine des technologies de surveillance de masse déployées dans le cadre des politiques post-9/11 (Musiani, 2015 ; Chateauraynaud, 2013).

Dans cet article, nous proposons d'interroger la qualité de l'alerte formulée par Kosinski et Wang et, par là, de questionner la position de « lanceur d'alerte » que les auteurs assument explicitement[5]. C'est pourquoi nous essaierons d'évaluer la robustesse de l'alerte au regard de la controverse qu'elle a suscitée en examinant dans quelle mesure le déplacement du particulier au général s'avère pertinent. Notre démarche consistera à dissocier clairement les trois composantes de l'alerte (performance du *deep learning* dans la perception, naturalisation de l'orientation sexuelle et transposition à l'espace public) afin de rendre plus intelligibles leurs faiblesses respectives.

L'ALERTE AMBIGUË DE LA MENACE D'UNE « A.I. GAYDAR »

Michal Kosinski et l'alerte sur les risques pour la vie privée à l'ère du *big data*

Figure saillante dans le domaine des sciences comportementales et en particulier de la psychométrie exploitant le *big data* (Kosinski *et al*., 2016 ; Kosinski et Behrend, 2017), Michal Kosinski se présente comme « computational

---

that reduce the risks faced by homosexual communities across the world. » Notre traduction : « Nous n'avons pas créé un outil intrusif de la vie privée ; nous avons plutôt montré que des méthodes basiques et largement utilisées posent de graves menaces pour la vie privée. Nous espérons que nos découvertes informeront le public et les décideurs politiques et les inciteront à concevoir des technologies et à élaborer des politiques qui réduisent les risques encourus par les communautés homosexuelles du monde entier. » (Wang et Kosinski, 2018, p. 255). La référence aux politiques de Ramzan Kadyrov est tacite chez Kosinski et Wang, explicite dans la réception médiatique.

5. Chateauraynaud et Torny ont proposé (1999) le terme de « lanceur d'alerte » comme alternative au terme anglais « whistleblower » et aux concepts de « prophète de malheur » de Jonas ou « dénonciateur » de Boltanski. Le terme s'est répandu et a été adopté en langue française. Dans les termes proposés par Chateauraynaud (2013), lanceur d'alerte est « [t]oute personne, groupe ou institution qui, percevant les signes précurseurs d'un danger ou d'un risque, interpelle une ou plusieurs puissances d'action, dans le but d'éviter un enchaînement catastrophique, avant qu'il ne soit trop tard ».

social scientist ». Depuis la fin des années 2000, son programme de recherche s'est essentiellement concentré sur la prédiction des comportements sociaux déterminés par des dispositions biologiques innées, par le biais des traces numériques, tout en ayant recours à du *machine learning*. C'est dans le cadre de ce programme que Kosinski va s'intéresser à l'orientation sexuelle comme « trait psychodémographique intime » servant à tester de nouvelles approches computationnelles de la prédiction comportementale à partir de données obtenues sur Internet. Son intérêt spécifique pour l'orientation sexuelle n'est donc que marginal par rapport à ce qui constitue le cœur de sa démarche et plus encore de ses conclusions, à savoir les conséquences en termes de vie privée qui en découleraient[6].

Les recherches de Kosinski suscitent l'attention des médias depuis quelques années. C'est d'abord en tant que co-concepteur de *myPersonality*, projet de psychométrie à partir de traces numériques mené dès 2009 à l'Université de Cambridge, que Kosinski va connaître une médiatisation importante. Reposant sur une approche psychodémographique de plus de 8 millions de profils Facebook (Kosinski *et al*., 2015), *myPersonality* s'appuyait sur une application Facebook pionnière dans la proposition de tests de personnalité. En utilisant cette application, les utilisateurs transmettaient les données de leur profil, participant ainsi aux recherches de Kosinski et de son collègue David Stilwell. Ce projet donnera lieu à un grand nombre de publications à fort impact, aussi bien dans des revues prestigieuses de psychologie ou de sciences sociales que dans des *proceedings* de conférences en informatique.

C'est un article au titre particulièrement explicite qui sera à l'origine d'une première alerte : « Private traits and attributes are predictable from digital

6. Lors d'un débat sur Twitter, Kosinski s'en est d'ailleurs expliqué très clairement : « If this makes you feel any better, original drafts did not include any theoretical claims – but focused on privacy » (notre traduction : « Si cela vous rassure, les premières versions du papier n'incluaient aucune affirmation théorique – elles étaient axées sur la vie privée ») (@michalkosinski en réponse à @thomas_thinks, le 9 septembre 2017 à 5 h 22, https://twitter.com/michalkosinski/status/906538340157095936) et « The main role of this paper is not to study PHT (there are better ways) but to draw attention to privacy risks » (notre traduction : « Le principal rôle de cet article n'est pas d'étudier la PHT [il existe de meilleurs moyens], mais d'attirer l'attention sur les risques pour la vie privée »), (@michalkosinski en réponse à @thomas_thinks, le 9 septembre 2017 à 5 h 24 https://twitter.com/michalkosinski/status/906538775551016960), consulté le 05/10/2018. Voir aussi les propos de Kosinski récemment recueillis par *The Guardian* à propos de sa conception déterministe et naturalisante des comportements sociaux (Lewis, 2018).

records of human behavior »[7] (Kosinski, Stillwell et Graepel, 2013). Suscitant un vif débat médiatique, cette recherche est réputée avoir initié chez Facebook des mesures en matière de protection des données dès 2014[8] (Grassegger et Krogerus, 2017 ; Murphy 2017). Depuis lors, Kosinski deviendra un personnage médiatique, tantôt invité pour parler de sa capacité à prédire la personnalité des individus, tantôt en tant qu'expert susceptible d'imaginer, du fait de ses propres études, ce que les grandes entreprises du numérique ou les États sont en mesure de dire et de prédire sur les individus.

C'est en 2017 que Kosinski va connaître le paroxysme de sa médiatisation. Dans le sillon de l'élection de Donald Trump, *myPersonality* est alors présenté comme une influence de la démarche de *Cambridge Analytica*, dont les méthodes controversées de communication politique ciblée auraient joué un rôle clé lors de la campagne électorale de Donald Trump fin 2016 (Grassegger et Krogerus, 2017). L'ambiguïté de Kosinski s'est alors retrouvée médiatisée : entendu à la fois comme celui qui dénonce l'état de fait et celui qui est suspecté d'avoir participé à son avènement.

L'alerte face à la critique

Entre résurgence de la physiognomonie par le biais du *deep learning* et affirmation de la détermination prénatale de l'orientation sexuelle, la nouvelle alerte de Kosinski connaîtra une réception agitée. *The Economist*, premier média à avoir donné de la visibilité à cette recherche, donne le ton en la présentant dans un dossier titré « Nowhere to hide : What machines can tell from your face »[9] (*The Economist*, 2017a, 2017b). Nombre de publications suivront dans la presse généraliste, avec des titres sensationnalistes sur la menace d'une « A.I. gaydar » susceptible de dire « si vous êtes homosexuel ». Reprenant les conclusions de Kosinski et Wang sans en examiner la validité ou la portée, ces premières publications poseront les bases d'un débat portant essentiellement

7. Notre traduction : « Les traits et attributs privés peuvent être prédits à partir des traces numériques du comportement humain ».
8. Ces mesures ainsi que la relation à *myPersonality* sont explicites dans les réponses de Facebook adressées le 8 juin 2018 au *Senate Committee on the Judiciary and Senate Committee on Commerce, Science, and Transportation*, à la suite de l'audition, pages 8, 9, 121, 145, 166, 207-208 et 216 (Facebook Inc., 2018), https://www.judiciary.senate.gov/imo/media/doc/Zuckerberg%20Responses%20to%20Judiciary%20Committee%20QFRs.pdf, consulté le 05/10/2018.
9. Notre traduction : « Nulle part où se cacher : ce que les machines peuvent dire à partir de votre visage ».

sur les enjeux éthiques de la recherche en I.A. et les implications pour les personnes LGBT (Hawkins, 2017 ; Levin, 2017b ; Libération, 2017).

Prenant appui sur le pre-print et les comptes rendus de presse, le débat s'est rapidement déployé dans la presse écrite en ligne et les médias sociaux, engageant des journalistes, des associations et personnes LGBTQ, des chercheurs et des praticiens de domaines très divers. Kosinski y participera lui-même, revendiquant son rôle de lanceur d'alerte et d'allié de la communauté LGBTQ. Or, si certains forums de la communauté LGBTQ, dont LGBTQ Nation, ont accueilli favorablement l'alerte contre ce qui serait une menace potentielle (Bollinger, 2017), d'importantes associations internationales comme *Human Rights Campaign* (HRC) et *Gay and Lesbian Advocates & Defenders* (GLAAD) ont manifesté un rejet catégorique de l'étude, de ses conclusions et de l'alerte, exhortant Stanford à « se distancier de cette science de pacotille (*junk science*) » (Anderson, 2017).

Face à la montée des critiques publiques, Kosinski et Wang ont rapidement réagi par la publication de deux documents sur Google Docs, un communiqué de réponse à la GLAAD et la HRC (Kosinski et Wang, 2017a) et une « Authors' note » dans laquelle ils reprennent point par point des objections reçues (Kosinski et Wang, 2017b). Selon les auteurs, le rejet de l'étude par ces associations était infondé, prématuré et irresponsable vis-à-vis des personnes LGBTQ, car elle détournerait l'implication principale de leur recherche, à savoir que la technologie était facilement disponible et pouvait dès lors être utilisée. Et Kosinski et Wang de conclure : « If our findings are wrong, we merely raised a false alarm »[10] (Kosinski et Wang, 2017a). Or, précisent-ils, le seul moyen de les démentir serait par des données scientifiques et par l'exercice de la réplication, et non par des personnes mal documentées, certes bien intentionnées, mais sans formation scientifique.

Déconstruire l'alerte sur la machine prédictive de Kosinski

Les réactions de Kosinski à la réception de l'article posent ainsi la question des conditions de la critique de ce type de recherche, et constituent en elles-mêmes un défi pour les sciences sociales. Comment un examen critique de l'alerte et de leur protocole peut-il être mené, compte tenu de la relative

10. Notre traduction : « Si nos conclusions sont fausses, nous avons simplement déclenché une fausse alerte ».

opacité dissuasive propre aux algorithmes de *deep learnig*, et sans disposer des moyens nécessaires à la réplication scientifique[11] ?

La critique, nous le montrerons, prend deux directions : l'une, prédominante, axée sur le rejet *in limine* de l'étude, éludant l'examen du lien entre la méthode de *deep learning*, l'interprétation des résultats et la justification de l'expérience par l'alerte ; l'autre, inspectant au contraire les écarts entre le protocole et les conclusions de l'article. Les critiques portant sur le protocole vont alors avancer l'hypothèse selon laquelle, compte tenu de l'opacité du réseau de neurones, il serait plus raisonnable de se pencher sur l'hypothèse que le modèle s'appuierait seulement sur la différence de styles d'autoprésentation entre homosexuels et hétérosexuels, y compris dans la manière de se prendre en photo, sans avoir nécessairement recours à la structure du visage[12]. La question qui se pose alors est précisément de savoir comment étayer cette hypothèse sans avoir accès à leurs données.

La controverse autour d'une « A.I. gaydar » est en ce sens particulièrement éclairante pour examiner l'entrée précipitée des algorithmes prédictifs dans nos existences et dans le débat public, tout en soulignant les enjeux qu'elle soulève pour les sciences sociales (Benbouzid, 2017 ; Cardon, 2015, 2018 ; Ziewitz, 2016). Si les algorithmes apparaissent comme des entités invasives dont le « pouvoir social » ne cesse de croître (Beer, 2017), leur opacité les fait apparaître comme des « boîtes noires » impénétrables, difficiles à scruter et, *in fine*, à critiquer (Diakopoulos, 2014). « Ouvrir les boîtes noires », injonction qui a historiquement guidé la sociologie des sciences et des techniques, est devenu l'un des grands défis de l'étude et de la gouvernance des algorithmes. Or, à la suite de Burell (2016), nous estimons que les choses techniques sont entourées de plusieurs niveaux et de plusieurs types d'opacité, qu'il n'y a pas strictement un intérieur et un extérieur de la « boîte noire », et que l'accès au code source et aux *datasets* n'est pas toujours indispensable pour démontrer l'inadéquation entre des résultats et leurs interprétations.

---

11. La première version du wiki proposée par la plateforme OSF permet de constater que de nombreuses informations, dont les photos, auraient été accessibles sous conditions pour les chercheurs qui en faisaient la demande. Cette possibilité a disparu dans les versions plus récentes.

12. Kosinski défend néanmoins l'idée selon laquelle l'autoprésentation ne serait pas contradictoire avec la PHT, qui la déterminerait tout autant que l'orientation sexuelle (Kosinski et Wang, 2017b).

La question se pose notamment pour les chercheurs en sciences sociales, pour lesquelles les « boîtes noires » ne devraient pas être dissuasives, car il n'y a pas nécessairement besoin de les ouvrir pour critiquer l'expérience. Notre enquête montre en effet comment l'examen des propositions de la recherche de Kosinski et Wang ainsi que de leur réception suffit à émettre une critique pour questionner la pertinence des conclusions et, partant, de l'alerte supposée. Leur « boîte noire » se résume, nous le montrerons, à l'usage circonscrit du *deep learning* à la détection des caractéristiques du visage (étude 1a) et aux *datasets* que les auteurs ont utilisés. Tout le reste de leur démarche est très conventionnel, dont le modèle prédictif, qui repose sur une simple régression logistique.

Afin d'établir les relations entre le protocole de Kosinski et Wang et les conclusions de leur article, d'une part, et entre ces conclusions et sa critique publique d'autre part, nous avons analysé avec précision le compte rendu que les auteurs font de l'expérience. Nous avons aussi identifié les réactions publiques à l'étude de Kosinski et Wang afin de mieux saisir les points d'appui et les brèches à partir desquels la critique s'instruit[13]. Par cette analyse en deux temps, nous livrons les résultats d'une enquête susceptible de rendre plus intelligibles la fabrication et la trajectoire possible d'une machine prédictive en société. Au terme de cette analyse, nous aurons montré la vacuité de l'alerte telle que formulée par les auteurs sur la base de leur recherche : la fin de la vie privée face à la capacité du *deep learning* à percevoir des dispositions morphologiques inaltérables, prédictives de l'orientation sexuelle et pourtant imperceptibles par des humains.

13. Le corpus que nous avons constitué pour l'analyse comprend deux ensembles. D'une part : le pre-print de l'article de Kosinski et Wang (dans l'ensemble de ses versions disponibles en ligne), la version finale publiée en février 2018, les documents publiés par les auteurs pour préciser leur propos (« Authors' note », communiqué de presse), et les posts et réponses de Kosinski à ses critiques sur Twitter. D'autre part : la quasi-totalité des publications en ligne mentionnant l'article ou plus généralement les recherches de Kosinski. Ce second ensemble comprend des articles de presse généraliste et spécialisée (anglo-saxonne et française), les documents de prise de position des associations LGBT, et les publications de chercheurs de différentes origines disciplinaires et de praticiens du *machine learning* sur des sites web et des médias sociaux jusqu'en février 2018.

À LA SOURCE DES ÉNONCÉS : RETOUR SUR LE COMPTE RENDU DE L'EXPÉRIENCE

L'article de Kosinski et Wang présente un compte rendu de leur expérience qui s'organise en 5 études, dont ils présentent les méthodes et les résultats respectifs. Chaque étude aboutit à la production d'un énoncé, qui participe de la démonstration générale :

1. grâce à un *réseau de neurones*, il est possible de produire un modèle capable de prédire l'orientation sexuelle d'un individu à partir d'images faciales issues d'un site de rencontres, avec une très grande précision, toujours supérieure au hasard ;
2. les visages des homosexuels présentent un caractère *atypique en termes de genre*, ce qui conforterait la théorie de l'exposition prénatale aux hormones (PHT) ;
3. la *morphologie du visage* à elle seule est prédictive de l'orientation sexuelle, conformément aux prédictions de la PHT ;
4. la plus faible performance de *juges humains* est conforme à celles des études antérieures, indiquant que le corpus d'images faciales issues d'un site de rencontre n'est pas particulièrement révélateur de l'orientation sexuelle ;
5. la performance du modèle à partir d'images présentes sur Facebook est comparable, signifiant que le *modèle prédictif* ne dépend pas de la source des images utilisées.

Dans cette section, il s'agit d'expliciter les protocoles et les résultats à partir desquels Kosinski et Wang tirent leurs conclusions, dont l'alerte sur la fin de la vie privée.

L'usage du *deep learning* pour prédire l'orientation sexuelle à partir d'images faciales issues d'un site de rencontres

Dans l'étude 1a, qui constitue le cœur de leur article, Kosinski et Wang travaillent à démontrer, en utilisant un algorithme de reconnaissance faciale et un classificateur binaire, qu'il est possible de prédire l'orientation sexuelle à partir d'images faciales obtenues d'un site de rencontre.

Constituer un corpus de données à partir d'un site de rencontres

La première étape de l'étude, c'est-à-dire le choix de la source, le nettoyage et la préparation des données, mérite une attention particulière. Dans la continuité des travaux de Kosinski sur la prédiction des « traits intimes » à partir de traces numériques (par ex., Kosinski, Stillwell et Graepel, 2013), les auteurs ont recours aux photos publiées sur un site de rencontres basé aux États-Unis. Suivant l'approche adoptée dans une série d'études sur le « gaydar » humain (par ex., Rule et Ambady, 2008), Kosinski et Wang se positionnent en porte-à-faux vis-à-vis des évaluations antérieures qui ont recours à des photographies prises en laboratoire, notamment pour des raisons de standardisation des expressions et des postures (par ex., Skorska *et al*., 2015). Par rapport aux photographies prises en laboratoire, celles qui sont déposées spontanément sur des sites web présentent l'avantage de constituer une base de données massive, à faible coût, tout en évitant les biais inhérents à la prise d'images en laboratoire. De surcroît, le site de rencontre fournit des informations autodéclarées, dont l'âge, la localisation et la préférence sexuelle. L'utilisation d'images autopostées participerait ainsi à la validité écologique de l'étude.

Sans donner de précision quant au nom du site ou à la méthode d'obtention des images, ils vont ainsi constituer un premier ensemble de 301 101 images issues de 75 223 profils publics de femmes et d'hommes ayant déclaré être basés aux États-Unis et avoir entre 18 et 40 ans, en veillant à ce qu'il y ait autant de profils homosexuels qu'hétérosexuels dans leur *dataset*, en considérant l'information relative au sexe des partenaires souhaités, telle que déclarée sur les profils.

Le recours à une telle source d'images n'est cependant pas sans inconvénient. Par exemple, les images faciales autopostées en ligne présentent des variations importantes en qualité, ainsi que sur le plan des expressions faciales, de l'orientation de la tête ou encore de l'arrière-plan. Afin de composer un ensemble d'images conformes aux exigences de leur expérience, Kosinski et Wang vont dès lors chercher à resserrer leur corpus autour de photographies décrivant exclusivement des visages comparables entre eux, en se limitant à des images faciales d'individus caucasiens qui respectent des critères précis de composition. Faisant appel aux services de la plateforme de *crowdsourcing* Amazon Mechanical Turk, ils demandent à des « microtravailleurs » (dits Turkers) de vérifier que les visages retenus représentent bien des adultes caucasiens dont les visages sont entièrement visibles et d'un sexe correspondant à celui indiqué sur le profil de l'utilisateur.

En complément de cette sélection humaine, les auteurs utilisent Face++[14], un service de *cloud computing* pour l'analyse faciale, afin d'extraire des informations sur la position du visage, les traits du visage, l'inclinaison, la rotation et le basculement de la tête, puis pour évincer toute photographie contenant plusieurs visages, contenant des visages partiels, trop petits ou trop inclinés ou dont le basculement ou la rotation sont trop élevés (10 et 15° respectivement). Kosinski et Wang cherchent ainsi à s'assurer que les différences apparentes de physionomie ne s'expliquent pas par la position de l'appareil photographique vis-à-vis du visage, mais bien par le visage lui-même.

Après une égalisation finale par genre et orientation sexuelle, le corpus utilisé dans l'étude contiendra 35 326 images (avec une proportion de 50/50 pour l'orientation sexuelle et de 53/47 pour le genre) et 14 776 profils uniques (avec des variations d'une à cinq images par profil). Suivant une procédure classique de validation croisée, le corpus sera aléatoirement divisé en 20 sous-corpus de tailles égales, dont 19 d'entraînement et 1 de test. Les données à partir desquelles les résultats seront présentés, le corpus de test, n'auront ainsi pas servi à entraîner le modèle.

Extraire les données faciales des images avec l'algorithme VGG-Face

Kosinski et Wang vont ainsi procéder à l'extraction de données descriptives qui serviront par la suite à la construction du modèle prédictif. Ils utilisent pour cela VGG-Face, un réseau neuronal profond de reconnaissance faciale. Mis au point par un groupe de l'Oxford Vision Lab de l'Université d'Oxford et librement disponible[15], il s'agit d'un algorithme appartenant à la famille des descripteurs à base de réseau de neurones convolutif (CNN) et qui sont à l'origine des progrès récents dans le domaine de la vision par ordinateur. VGG-Face se positionne dans un espace dominé par Google et Facebook, face aux algorithmes desquels VGG-Face témoigne de performances compétitives, obtenues à partir de bases de données pourtant moins volumineuses (Parkhi, Vedaldi et Zisserman, 2015). Kosinski et Wang soulignent que ce CNN est particulièrement adapté à leur recherche morphopsychologique, puisqu'il est entraîné afin de maximiser les traits stables d'un visage et minimiser les traits

14. Face++ est proposé par l'entreprise chinoise Megvii. Les auteurs ont eu accès à Face++ à titre gracieux. La présentation de l'outil peut être consultée à http://www.faceplusplus.com, consulté le 05/10/2018.

15. Les documents relatifs au descripteur ainsi que le code source sont disponibles sur la page : http://www.robots.ox.ac.uk/~vgg/software/vgg_face/, consulté le 05/10/2018.

les plus variables (le maquillage par exemple, mais aussi les effets de posture ou de position du visage vis-à-vis de l'objectif), son objectif initial étant de reconnaître des personnes dans des configurations différentes.

Entraîner un algorithme de classification binaire de l'orientation sexuelle

Une fois l'extraction des caractéristiques faciales terminée (4 096 scores par image), Wang et Kosinski s'attellent à associer, pour chaque individu, la caractérisation des images avec l'état binaire de leurs préférences sexuelles : homosexuelles ou hétérosexuelles. Les chercheurs utilisent alors une régression logistique. Modèle éprouvé en *machine learning*, la régression logistique pondère chaque variable quantitative en entrée (ou *features*) en fonction de l'importance qu'elle possède dans la détermination de la sortie binaire, telle qu'être élu ou ne pas être élu, tomber malade ou ne pas tomber malade, etc. Kosinski et Wang génèrent ainsi un classificateur binaire de l'orientation sexuelle à partir des 19 sous-corpus d'entraînement.

Tester et évaluer la performance du modèle prédictif de l'orientation sexuelle

Kosinski et Wang vont alors tester leur modèle prédictif de l'orientation sexuelle en l'appliquant à des images qui n'ont pas été utilisées lors de la phase d'entraînement. Chaque image de cet échantillon se voit attribuer une probabilité de décrire un individu homosexuel. Pour évaluer la robustesse du modèle, les auteurs choisissent d'utiliser l'une des méthodes d'évaluation des classificateurs binaires les plus courantes : l'AUC (*area under curve*). Il s'agit de la mesure de la surface en dessous de la courbe sensibilité/spécificité, appelée courbe ROC (*receiver operating characteristic*). Cette méthode d'évaluation consiste à classer les individus selon leur probabilité d'être homosexuel, puis à représenter la relation entre le taux de vrais positifs et de faux positifs, à mesure que l'on considère un nombre de plus en plus important d'individus[16]. L'AUC est interprétée dans le contexte de cette étude comme la probabilité que le classificateur discrimine correctement, lorsqu'il est soumis à une image d'une personne homosexuelle et une image d'une personne hétérosexuelle, sélectionnées aléatoirement parmi l'ensemble des images. Une AUC de 0,5

16. Pour plus de détails sur l'interprétation probabiliste de l'AUC, voir Hanley et McNeil (1982). Sur les choix de conception des expériences en sciences sociales computationnelles et les effets sur l'interprétation des performances découlant du recours à un ou un autre standard d'évaluation des modèles prédictifs, dont l'AUC, voir notamment Hofman, Sharma et Watts (2017).

signifie en cela que le classificateur a une chance sur deux de se tromper, ce qui correspond à une performance équivalente au hasard. Un classificateur parfait disposerait ainsi d'une AUC de 1. Le modèle proposé par Kosinski et Wang obtient une AUC variant de 0,81 à 0,91 dans la détermination de l'homosexualité chez les hommes, et de 0,71 à 0,83 chez les femmes (selon le nombre d'images disponibles, de 1 à plus de 5). Cette performance est très élevée dans le cas des hommes disposant de 5 images (0,91), mais il est important de la circonscrire à l'interprétation spécifique de cet indicateur : il ne s'agit aucunement de la probabilité de classifier correctement une image, mais de classifier correctement une image parmi deux images, dont l'une représente une personne homosexuelle et l'autre une personne hétérosexuelle. Cette nuance est subtile, mais très importante, puisqu'elle est au cœur de la fragilité de l'alerte que les auteurs poseront en conclusion de leur recherche.

Expliquer la performance du modèle conformément à la PHT

Afin de conforter la relation entre la performance du modèle et la PHT, Kosinski et Wang vont chercher à identifier les éléments des images qui sont déterminants pour la classification (étude 1b). Observant que ce n'est pas le fond des images qui sert le modèle, ils chercheront alors les caractéristiques du visage – ou plutôt de l'image du visage – qui sont les plus déterminantes dans la prédiction de l'orientation sexuelle (étude 1c). Les études suivantes (2 et 3) vont alors tester l'hypothèse du lien entre la PHT et la capacité prédictive du modèle.

Identifier les zones de l'image les plus déterminantes pour le prédicteur

Les auteurs vont chercher à vérifier si le modèle de régression logistique a bien considéré les traits stables des visages et non leur environnement (étude 1b). Afin de déterminer les portions de l'image qui ont compté dans la détermination du modèle prédictif, les auteurs vont tester de nouveau le modèle sur deux sets d'images (chacun de 100 images, l'un d'hommes, l'autre de femmes) auxquelles seront masqués itérativement des bouts de 7 par 7 pixels, et évaluer la chute de performance du modèle afin d'identifier les zones les plus exploitées. Les résultats indiquent que la régression logistique s'appuie effectivement sur des éléments du visage (le nez, les yeux, les sourcils, les joues, la racine des cheveux et le menton chez les hommes, et le nez, les coins de la bouche, les cheveux et le tour de cou chez les femmes) et non sur le fond des photographies.

Examiner les traits faciaux servant à la prédiction de l'orientation sexuelle

Kosinski et Wang vont alors considérer les images faciales avec les plus hautes et les plus basses probabilités d'être homosexuel pour produire quatre images composites reflétant ce que seraient des visages typiques de femmes et d'hommes homosexuels et hétérosexuels (étude 1c). L'interprétation « à l'œil nu » des images ainsi générées laisse suggérer aux auteurs des résultats conformes à la PHT. Il semblerait que les hommes hétérosexuels et les femmes lesbiennes aient des mâchoires plus larges que les hommes gays et les femmes hétérosexuelles. À ces différences morphologiques s'ajouteraient des différences au niveau de la capillarité, du teint et du style, y compris le port apparent de casquettes de base-ball chez les hommes hétérosexuels et les femmes lesbiennes. Les auteurs laissent ainsi entendre que les homosexuels seraient globalement atypiques du point de vue du genre (*gender atypical*), suggérant des femmes plus masculines et des hommes plus féminins que la moyenne.

Tester l'hypothèse de l'atypicité de genre des visages homosexuels

Afin de tester cette hypothèse, ils entreprennent de mesurer la féminité faciale des images utilisées (étude 2). Ils utilisent pour cela un modèle prédictif du genre basé sur un autre réseau neuronal issu de *myPersonality* (Kosinski *et al*., 2015). Si les scores de féminité décrivent le genre des individus avec un AUC de 0,98, le croisement entre les scores de féminité et l'orientation sexuelle déclarée produit en revanche une AUC de seulement 0,58 pour les femmes et de 0,57 pour les hommes, ce qui ne dispense pas les auteurs de suggérer que la féminité faciale permet de discriminer les visages gays et hétérosexuels avec « une certaine précision » (*some accuracy*). Considérant par ailleurs que la mesure de la féminité des visages d'hommes montre une corrélation positive avec la probabilité d'être gay et que l'inverse est observé pour les femmes, Kosinski et Wang en concluent que les personnes homosexuelles ont un visage atypique selon le genre.

Déterminer le rôle prédictif de la morphologie faciale

Kosinski et Wang tentent alors d'évaluer le rôle spécifique de la morphologie dans la prédiction de l'orientation sexuelle (étude 3). VGG-Face extrait en effet des informations relatives à la morphologie du visage. Wang et Kosinski utilisent dès lors 83 repères extraits par Face++ décrivant la forme des sourcils, des yeux, du nez, de la bouche et de la mâchoire. Pour chaque élément, la distance entre les repères est calculée. Ces distances viennent alors se substituer

aux scores produits par VGG-Face. Une nouvelle régression logistique génère alors un modèle dont la performance est à peine inférieure (AUC de 0,85 pour les hommes et de 0,75 pour les femmes). La forme de la mâchoire semble de surcroît être l'indicateur le plus prédictif, tant pour les hommes que pour les femmes, alors même que la mâchoire ne serait pas affectée par le maquillage. La morphologie, seule, semble ainsi prédictive de l'orientation sexuelle. La forte diminution de l'information utilisée n'entraîne en effet qu'une légère diminution des AUC (passage des 4 096 scores de VGG-Face à de simples distances euclidiennes calculées à partir de quelques repères de Face++).

Généraliser les découvertes au-delà du site de rencontres

Les conclusions des études présentées par les auteurs suggèrent que le modèle peut prédire l'orientation sexuelle à partir d'images d'un site de rencontre, en s'appuyant seulement sur les différences physionomiques ou même morphologiques des visages. Sur cette base, Kosinski et Wang vont alors chercher à montrer que leurs conclusions peuvent être généralisées au-delà de ce site de rencontres en s'assurant que ces images ne sont pas particulièrement suggestives de l'orientation sexuelle (étude 4 et 5).

Démontrer que les images utilisées ne sont pas spécialement révélatrices de l'orientation sexuelle pour des humains

Dans l'étude 4, Kosinski et Wang reviennent alors sur des risques qui ne relèvent pas de l'*overfitting* (surajustement du modèle aux données), mais plutôt du biais de confirmation (surajustement des données à l'hypothèse). Il se peut en effet que les images postées sur un site de rencontres soient trop explicites, et qu'elles surexpriment les traits susceptibles de jouer dans l'orientation sexuelle – étant donné la visée de séduction inhérente aux images utilisées dans ces circonstances. L'expérience fait alors appel au classificateur de référence : l'humain. Les auteurs se réfèrent à des expériences précédentes où des humains avaient eu à prédire l'homosexualité sur des photographies prises en laboratoire ou dans des sites de rencontres (Ambady, Hallahan et Conner, 1999 ; Rule, Macrae et Ambady, 2009). L'AUC se situait alors entre 0,55 et 0,65. Pour tester un éventuel biais de surexpression de leur *dataset*, Wang et Kosinski font de nouveau appel à des Turkers. L'AUC des humains – dans un contexte artificiel de discrimination entre une personne hétérosexuelle et une personne homosexuelle – est de 0,61 pour les hommes et 0,54 pour les femmes. Les auteurs en concluent que leur *dataset* ne surexprime pas

l'orientation sexuelle, ou, du moins, telle qu'elle est susceptible d'être perçue par des humains.

Tester le classificateur sur des images issues de Facebook

La cinquième et dernière étude évalue la portée du modèle proposé au-delà des images des sites de rencontre. Il s'agit de tester le classificateur sur des photographies qui proviennent d'un environnement différent : Facebook. Les auteurs exploitent alors la base de *myPersonality* dont ils extraient 14 338 images de 6 075 hommes explicitement homosexuels. Afin d'accroître la certitude sur l'orientation sexuelle effective des hommes (sur lesquels ils ont choisi de se focaliser pour des raisons d'accessibilité à l'information comparativement aux femmes homosexuelles), Kosinski et Wang utilisent le *Facebook Audience Insight* pour identifier les 50 pages Facebook les plus populaires parmi les hommes gays. Sont ensuite identifiés les profils d'hommes qui ont déclaré leur intérêt pour au moins deux de ces pages (sous la forme du *like*) et être explicitement intéressés par d'autres hommes. Seuls les profils qui respectent ces deux conditions sont retenus.

Afin de tester la relative neutralité de la base constituée à partir du site de rencontre, un échantillon de cette base est confronté à celui constitué à partir de Facebook. Dans ce cas, une régression logistique ne parvient pas à prédire à quelle source appartient le profil (AUC de 0,53). Le classificateur réussit toutefois à distinguer les hommes gays issus de Facebook des hommes hétérosexuels issus du site de rencontre (AUC de 0,74). En dépit de la chute de performance du classificateur (de 0,81 à 0,74), les auteurs en concluent que la performance du classificateur ne dépend donc pas de la source des photographies utilisées et que, par conséquent, le modèle testé dans l'article est généralisable à toutes sortes de situations.

Conclusion et alerte sur les risques pour la vie privée

C'est ainsi que Kosinski et Wang en viennent à conclure que les 5 études présentées confirment l'hypothèse initiale (les visages comportent plus d'informations sur l'orientation sexuelle que les humains ne peuvent en percevoir) et confortent les prédictions de la PHT[17].

---

17. Il est important de noter que dans le pre-print mis en ligne début septembre les auteurs affirmaient que leurs résultats apportaient un « soutien fort » (*strong support*) à la PHT. Suite

Kosinski et Wang concèdent cependant qu'il ne faut pas pour autant surinterpréter les résultats de cette recherche. Ils s'arrêtent sur deux questions. Tout d'abord, ils rappellent qu'il ne s'agit que de probabilités. Parmi les individus se déclarant homosexuels, rien n'empêche des femmes d'être « féminines », et des hommes d'être « masculins ». Ils précisent ensuite la manière dont il convient d'interpréter la performance importante du modèle pour les visages d'hommes disposant de 5 images ou plus. À cet égard – et c'est là un point fondamental –, Kosinski et Wang soulignent que l'AUC de 0,91 ne signifie pas que 91 % des gays peuvent être identifiés ou que la classification est correcte dans 91 % des cas. Ainsi, Kosinski et Wang expliquent que lorsqu'on considère l'AUC de 0,91 en tentant compte de la prévalence estimée de l'homosexualité masculine aux États-Unis de 7 %, sur les 70 images d'hommes sur les 1 000 disposant de la plus haute probabilité de décrire un homme gay selon le modèle, 39 le seraient vraiment[18]. La performance serait ainsi 8 fois supérieure au hasard.

Kosinski et Wang en viennent à ce qui serait, selon eux, l'enjeu le plus critique de leur découverte, à savoir la fin de la vie privée. Les gouvernements et les entreprises, avertissent-ils, utiliseraient déjà des technologies sophistiquées pour déduire les traits intimes des citoyens. Ainsi, Kosinski et Wang se positionnent en lanceurs d'alerte, faisant valoir que leur recherche, loin d'encourager la mise en œuvre de cette machine prédictive, vise au contraire à montrer à quel point sa mise en œuvre serait simple dans d'autres contextes que la recherche. Selon eux, les caméras de vidéosurveillance et les images disponibles en abondance sur Internet pourraient même alimenter des classificateurs encore plus précis. Ils concluent dès lors leur démonstration en insistant sur le fait que l'érosion de la vie privée semblant inévitable, seule l'éducation à la tolérance permettra un espace pacifié et accueillant.

LES ÉNONCÉS À L'ÉPREUVE DE L'ESPACE PUBLIC : UNE ÉVALUATION SAUVAGE

En articulant des sujets sensibles et controversés, et en concluant par un enjeu majeur de société, les propositions de Kosinski et Wang se sont exposées

---

aux critiques adressées après la publication, les auteurs ont nuancé ce propos, remplaçant la formulation « soutien fort à » par « cohérentes avec » (*consistent with*) la PHT.

18. Les auteurs ont procédé à une simulation en reconstituant aléatoirement un échantillon représentatif de cette répartition afin d'appuyer leur propos.

à de nombreuses critiques. À la suite des articles de *The Economist* (The Economist, 2017b), *The Guardian* (Levin, 2017a, 2017b), ou encore du *New York Times* (Kuang, 2017 ; Murphy, 2017), les critiques se sont déployées dans l'espace public à partir de septembre 2017 et marquèrent le début d'une controverse. Si les critiques sont essentiellement informées par la lecture des comptes rendus de presse (Bjork-James, 2017), d'autres se sont confrontées aux différentes études réalisées et permettent d'identifier les points les plus sensibles de cette recherche. C'est ainsi que le débat prendra la forme d'une « évaluation sauvage », du fait de la précipitation d'acteurs tant universitaires qu'extra-universitaires, et de la temporalité accélérée des réactions médiées par Internet.

Les critiques de l'étude se sont concentrés alternativement sur deux aspects de la recherche :

1. la contestation de ses buts, ses fondements conceptuels, ses hypothèses de départ, et des choix concernant le corpus ;
2. les méthodes, les mesures et les raisonnements, tels que présentés dans le compte rendu, et à partir desquels les revendications faites par les auteurs peuvent être étayées.

L'ensemble de ces critiques permet de souligner les limites méthodologiques de la démarche, mais aussi le décalage entre la performance du modèle et les enjeux qui lui sont associés. Dans cette section, nous mettons ces critiques en perspective, avant de nous concentrer sur le travail de totalisation de celles portant sur l'articulation des protocoles et des conclusions que les auteurs en tirent.

Rejet des choix fondamentaux à la base de l'étude

Les premières critiques à avoir gagné une notoriété publique ont été celles avancées par les représentants des associations LGBTQ. Du point de vue de la HRC (Human Rights Campaign) et de la GLAAD (Gay & Lesbian Alliance Against Defamation), les résultats seraient erronés, la méthodologie biaisée et le *peer review* manquant. Les médias auraient été piégés par un argumentaire trompeur qui suggère qu'une intelligence artificielle pourrait détecter l'orientation sexuelle. Loin d'identifier l'orientation sexuelle d'une personne, la technologie algorithmique en question se limiterait à

« reconnaître un modèle dans un petit sous-ensemble de personnes sur les sites de rencontres de lesbiennes et de gays blancs qui se ressemblent » (Anderson, 2017). En « imaginant un instant les conséquences potentielles si cette recherche biaisée était utilisée pour soutenir les efforts d'un régime brutal pour identifier et/ou persécuter les gens qu'ils croyaient être gays », ce serait bien la sécurité et la vie privée des personnes tant LGBTQ que non LGBTQ qui serait en péril, car elle nuirait « non seulement aux gays et lesbiennes qui se trouvent dans des situations où faire son *coming out* est dangereux, mais encore aux hétérosexuels erronément outés » (Anderson, 2017). Aussi, les représentants des associations LGBTQ exigeaient que les médias dénoncent l'étude et exhortaient Stanford à « se distancier de cette science de pacotille (*junk science*) » afin de ne pas « donner son nom et sa crédibilité à une recherche qui aggrave la situation du monde et le rend moins sûr qu'avant » (Anderson, 2017).

En ligne avec ces objections, la plupart de la critique émanant d'universitaires en sciences humaines et sociales et notamment en études de genre s'est concentrée, d'une part, sur le rejet de ce qui serait une étude de classification de personnes homosexuelles sur des bases physionomiques et, d'autre part, sur ses présupposés, manifestes notamment au niveau du corpus, des catégories binaires utilisées, et de l'interprétation « naturalisante » des résultats de l'expérience.

Les critiques ont d'abord porté sur les critères de sélection de la source des images et par conséquent de la population représentée (Mattson, 2017 ; Anderson, 2017). Le choix de limiter l'étude à un site de rencontres, et donc à la population relativement homogène de femmes et d'hommes blancs qui le fréquentent, exclut d'énormes segments de la communauté LGBTQ, y compris les personnes de couleur, les personnes transgenres, les personnes âgées et les autres personnes LGBTQ qui ne postent pas des photos sur les sites de rencontres. À ceci, la GLAAD ajoutait un manque de vérification des informations, y compris l'âge et l'orientation sexuelle déclarés en ligne. Les conclusions de l'étude ne porteraient ainsi que sur des normes de beauté sur les sites de rencontres (Anderson, 2017).

Les critiques se sont aussi adressées aux biais résultant d'une conception étriquée de la sexualité (Casilli, 2017 ; Mattson, 2017 ; Weber, 2017). La considération de départ, de ne retenir que deux orientations sexuelles – homosexuelle et hétérosexuelle – se répercute ainsi sur les résultats, avec l'exclusion des personnes bisexuelles et de toute une diversité de comportements, d'identités

non binaires et de désirs sexuels. La vision binaire, essentialiste et exclusive des orientations sexuelles humaines de l'étude révélerait l'hétéronormativité et les « préjugés anti-LGBT » des chercheurs de Stanford (Casilli, 2017), ce à quoi Kosinski et Wang ont répondu dans leur « Authors' note » que cela ne signifie pas qu'ils dénient l'existence des bisexuels et des personnes non binaires et que le modèle serait probablement plus performant s'ils avaient pu les considérer (Kosinski et Wang, 2017b).

À la suite de ces critiques, la recherche de Kosinski et Wang a fait l'objet d'un « examen éthique » supplémentaire (Flaherty, 2017). L'American Psychological Association, responsable du *Journal of Personality and Social Psychology*, le sollicita notamment dans le but de vérifier que le comité d'évaluation éthique de Stanford avait bien validé tous les aspects sensibles de la recherche. Les deux évaluations éthiques, celle de Stanford, puis celle de la prestigieuse revue qui l'a finalement publié, n'ont ainsi pas opposé d'arguments suffisants pour remettre en cause la recherche et sa publication.

Critique de l'interprétation des résultats à l'aune de la critique de la méthode et de l'algorithme

À ces critiques fondamentales sur les enjeux sociaux et éthiques de la recherche de Kosinski et Wang comme pratique scientifique inacceptable s'ajoutent d'autres critiques provenant le plus souvent de chercheurs et de praticiens en *machine learning* ou plus largement familiers des méthodes quantitatives en sciences sociales (Agüera y Arcas, Todorov et Mitchell, 2018 ; Bergstrom et West, 2017 ; Cohen, 2017 ; Howard, 2017). Cet ensemble de critiques porte plus spécifiquement sur l'articulation entre les conclusions et les interprétations, d'une part et, d'autre part, les résultats et les méthodes présentés dans l'article.

La critique face à l'opacité de la machine : la question de l'insensibilité de la sortie de VGG-Face aux caractéristiques faciales éphémères

Avant que l'apprentissage profond n'atteigne le niveau actuel, les chercheurs en *machine learning* devaient passer par une étape coûteuse et difficile de *feature engineering*. Celle-ci consiste à indiquer aux algorithmes quels *patterns* trouver, telle que la distance entre le sourcil et le nez, les lèvres et le menton, etc. Cette tâche est désormais automatisée et des progrès considérables afin d'être insensible à des éléments tels que les accessoires, le maquillage et la

pose. Or la procédure automatique, impliquant l'extraction, puis la réduction des caractéristiques de l'image, limite l'identification des éléments dont l'algorithme se sert effectivement pour déterminer l'orientation sexuelle. C'est pourquoi les critiques axées sur la méthodologie et la validité scientifiques des énoncés portent essentiellement sur l'identification des éléments dont dépend la capacité de l'algorithme à détecter l'orientation sexuelle.

Kosinski et Wang plaçaient le choix de VGG-Face au fondement du lien entre la PHT, la performance de leur algorithme et l'alerte. Afin de reconnaître une personne malgré des variations au niveau de l'aspect, l'expression et la pose, VGG-Face maximise les caractéristiques permanentes et structurelles des visages. Ce sont donc bien ces traits que VGG-Face aurait maximisés pour parvenir à la discrimination entre les deux orientations sexuelles (études 1b et 1c). Or la critique va insister sur le fait que VGG-Face n'en reste pas moins dépendant de la manière dont il est implémenté et du *dataset* sur lequel la version spécifique du logiciel a été entraînée.

L'argument de la critique se résume alors au constat suivant : dès lors que le logiciel n'est pas utilisé pour identifier les gens, mais pour trouver des modèles complexes entre différentes personnes comme le font Kosinski et Wang, VGG-Face ne serait plus insensible à l'expression et à la pose. C'est ce que Tom White, professeur de conception informatique à l'École de design de l'Université Victoria, Nouvelle-Zélande, a fait valoir en premier, se fondant sur une série d'expériences simples qu'il avait menées avec VGG-Face. Les résultats, partagés sur Twitter, montrent qu'un classificateur formé pour distinguer entre des visages heureux et des visages neutres pouvait atteindre une AUC de 0,92. Lorsque le classificateur était formé pour discriminer entre des visages tristes et des visages heureux, l'AUC augmentait à 0,96. Et lorsqu'il était demandé au classificateur de discriminer des images faciales selon la pose de la tête, il atteignait une précision de 100 %, classant correctement la totalité des images où la tête est tournée vers la droite sur un jeu de 576 images[19]. Ainsi, si les différences dans l'émotion et dans la pose pouvaient être prédites par ces modèles basés sur la sortie de VGG-Face, d'autres différences liées aux modes de présentation distinctifs des personnes homosexuelles et des personnes hétérosexuelles pouvaient être tout aussi décisives dans le modèle

19. Tom White, *@dribnet*, « Hopefully this firmly puts to rest the "vggface is invariant to pose" claims […] ». Notre traduction : « Espérons que ceci mettra définitivement un terme aux affirmations comme quoi "vggface est invariant à la pose" ». Tweet du 15 septembre 2017, https://twitter.com/dribnet/status/908521750425591808, consulté le 05/10/2018.

proposé par Kosinski et Wang. L'hypothèse serait plus raisonnable que celle axée sur les effets de la PHT. Les conclusions de White ont été par la suite reconduites maintes fois par la presse spécialisée (Gershgorn, 2017 ; Vincent, 2017), mais aussi dans les publications d'autres critiques, notamment celles de spécialistes en *machine learning* (Agüera y Arcas, Todorov et Mitchell, 2018 ; Howard, 2017).

Des performances prédictives et l'hypothèse du poids des différences d'autoprésentation

La contribution la plus significative visant à contrecarrer les explications « naturalisantes » de Kosinski et Wang est celle faite dans un billet de Blaise Agüera y Arcas et Margaret Mitchell, tous deux chercheurs spécialisés dans la reconnaissance faciale chez Google Photos à Seattle, avec Alex Todorov, professeur au département de psychologie de Princeton et directeur du laboratoire de perception sociale (Agüera y Arcas, Todorov et Mitchell, 2018).

Selon Agüera y Arcas et ses collègues (2018), les quatre visages composites d'hommes et de femmes homosexuels et hétérosexuels (étude 1b) révéleraient à l'œil nu beaucoup d'informations sur les éléments qui étaient disponibles à l'algorithme pour réaliser la classification et qui ne relèvent pas de la structure faciale : le sourire plus prononcé sur le visage composite caractérisant une femme hétérosexuelle comparativement à l'absence de sourire sur le visage caractérisant une femme lesbienne ; le fard à paupières moins perceptible dans l'image composite d'une femme lesbienne ; ou encore la marque autour des yeux suggérant un port de lunettes plus fréquent chez les homosexuels que chez les hétérosexuels. Ces observations suggéreraient qu'il est bien plus raisonnable d'estimer que les différences exploitées pour la prédiction soient liées au maquillage, au style et aux angles de prise de vue, c'est-à-dire à des traces de comportements sociaux présentes dans les photographies autopostées, et non pas à des différences morphologiques attribuables à la PHT. Pour tester cette hypothèse sans pourtant avoir recours aux données ou au même descripteur VGG-Face, Agüera y Arcas et ses collègues vont d'abord chercher à confirmer les caractéristiques observées via un questionnaire. En se servant des réponses, ils vont produire un simple classificateur de l'orientation sexuelle.

Agüera y Arcas et ses collègues ont mobilisé pour cela 8 000 Turkers américains auxquels ils ont proposé un questionnaire comportant 77 questions fermées portant tantôt sur le genre, l'identité et l'orientation sexuelle

(« avez-vous une attraction sexuelle pour le même sexe ? », « avez-vous une attraction romantique pour le même sexe ? » ou « êtes-vous gay ou lesbienne ? », etc.), tantôt sur la présentation et le style de vie (« portez-vous des lunettes ? », « aimez-vous comment vous vous voyez avec des lunettes ? », « avez-vous une barbe ? », etc.). La ventilation des réponses en fonction de l'âge du répondant et en considérant plusieurs catégories distinctes, tantôt axées sur l'attraction, tantôt sur l'identité, suggère que les personnes les plus jeunes sont beaucoup plus susceptibles de s'identifier de façon non hétéronormative. Les réponses aux questions sur le style de vie confirment par ailleurs les tendances observées sur l'image composite quant à l'ombre à paupières et le port de lunettes, ainsi que le port de la barbe chez les jeunes hommes attirés par le même sexe. Cela suggère, comme l'avait fait valoir Gregor Mattson (Mattson, 2017), l'un des spécialistes en étude de genre les plus actifs dans la contestation, que les tendances de la mode et les normes culturelles jouent un rôle dominant – des choix de style qui ne pourraient donc pas être attribués, selon Mattson, à la différence d'exposition prénatale aux hormones.

Se servant des réponses aux questionnaires, Agüera y Arcas et ses collègues ont ensuite produit un simple classificateur. Ils affirment obtenir des résultats comparables à ceux de Kosinski et Wang. Par exemple, pour deux femmes, dont l'une est lesbienne, l'autre hétérosexuelle, leur algorithme est précis à 63 % juste en considérant le port d'ombre à paupières. En ajoutant six autres questions fermées relatives à la présentation, la performance du modèle atteint 70 % (rien n'est précisé quant à la performance de cette démarche pour les hommes). La performance du prédicteur de Kosinski et Wang, concluent les trois chercheurs, serait ainsi probablement très influencée par la présence d'indices culturels distinctifs des personnes homosexuelles et hétérosexuelles.

L'ambiguïté entre l'inclinaison du visage et la dimension des traits mesurés

Afin d'infirmer l'hypothèse de l'importance des paramètres relatifs à l'autoprésentation, Kosinski et Wang ont montré que la mesure de repères morphologiques se révélait suffisante à la prédiction de l'orientation sexuelle. Ces repères pourraient néanmoins être eux aussi influencés par l'autoprésentation. La posture pourrait non seulement constituer une source de différenciation supplémentaire dans le cadre de l'étude 1 (*deep learning*), mais elle pourrait aussi expliquer les différences de contours faciaux identifiées dans les études 2 (genre atypique) et 3 (repères morphologiques).

Agüera y Arca et ses collègues suggèrent que, malgré les précautions prises par Kosinski et Wang pour standardiser les photographies autopostées, l'image faciale composite permet aisément d'observer des effets de prise de vue, c'est-à-dire des effets sur les proportions du visage de la position de l'appareil photo relativement au sujet photographié. Les hommes hétérosexuels prendraient plus souvent la photo en contre-plongée, les femmes hétérosexuelles tendraient à les prendre d'en haut, et les hommes gays, comme les femmes lesbiennes, tendraient à privilégier les prises de face. Une prise de vue en contre-plongée tend en effet à produire un élargissement du menton, une atténuation du sourire, un raccourcissement du nez, et un rétrécissement du front. Les hommes hétérosexuels, en exploitant la contre-plongée, essaieraient ainsi de produire des « effets de dominance ». Inversement, les femmes hétérosexuelles valoriseraient une prise de vue plus plongeante, qui agrandirait les yeux (effet qui pourrait être recherché afin d'apparaître plus « attrayantes »).

Ces critiques vont ainsi conclure qu'il serait plus vraisemblable que ce que Kosinski et Wang attribuent à la PHT soit dû à de simples effets résultant des choix de point de vue dans la prise du selfie.

La comparaison des performances des algorithmes et des humains dans la détection de l'orientation sexuelle

Les critiques se sont aussi portées sur l'étude 4, lors de laquelle Kosinski et Wang comparent la performance de leur modèle prédictif à celle des Turkers et en tirent trois conclusions : leur algorithme est plus à même de prédire l'orientation sexuelle que les humains ; il détecte des caractéristiques invisibles aux humains ; et l'échantillon d'images autopostées sur le site de rencontre n'est par conséquent pas particulièrement révélateur de l'orientation sexuelle.

À cela, Casilli (2017) a objecté que la composante humaine de la comparaison ne peut pas servir de référence pour faire valoir des revendications à portée universelle quant aux capacités des humains par rapport à celles d'un algorithme. Insistant sur le *digital labor* de la démarche, il questionne le recours à des « microtâcherons » payés quelques centimes de dollar, provenant uniquement des États-Unis, auxquels on demande de classer des images semi-standardisées de visages d'Américains, que d'autres Turkers américains ont estimé être caucasiens.

D'autre part, comme le notent d'abord Bergstrom et West (2017), puis Howard (2017), le problème de la comparaison entre ces Turkers et la machine est qu'elle confond leur capacité respective à accomplir deux tâches différentes :

détecter des indices dans les photographies et prendre de bonnes décisions sur la base de ces indices. En premier lieu, les machines semblent mieux équipées pour corriger leurs *a priori* face à la distribution réelle des faits et événements, alors que les humains seraient freinés par l'inertie de leurs *a priori* (Bergstrom et West 2017). Aussi, considérant les différences fondamentales entre l'humain et l'algorithme, Howard rappelle que lorsqu'on décide d'engager de telles comparaisons, la pratique recommandée exige de chercher à réduire les asymétries en donnant à l'humain la possibilité d'étudier les données d'apprentissage que l'ordinateur a reçues (Howard, 2017).

C'est pourquoi, en second lieu, Kosinski et Wang auraient dû donner à chaque Turker un grand nombre d'exemples de visages avec leurs étiquettes correspondantes (homosexuel ou hétérosexuel) avant de prédire l'orientation sexuelle des visages. Or Howard, ainsi que Bergstrom et West, suggèrent que la conclusion de Kosinski et Wang, qui supposent des caractéristiques physionomiques en dessous du seuil de détection humaine, constitue une violation du principe de la parcimonie : l'explication la plus parcimonieuse des résultats de l'étude 4 étant tout simplement que ce n'est pas un concours équitable.

Les performances du modèle sur la base des images issues de Facebook et du site de rencontres comme conditions de généralisation

Un autre ensemble de critiques questionne plus spécifiquement la généralisation de la performance du modèle à d'autres corpus. Sur la base de l'étude 5, Kosinski et Wang concluent que le classificateur de l'orientation sexuelle est insensible à la source des images. Cette conclusion est décisive pour étayer l'indépendance de la source et en cela la pertinence de l'alerte.

Le test de non-discrimination entre les images d'utilisateurs de Facebook gays et les images d'hommes gays du site de rencontres a soulevé des critiques qui ont fait observer que ce qui compte dans la considération de l'échec du classificateur à les distinguer (AUC de 0,53) n'est pas tellement l'environnement d'origine de l'image (Facebook *vs* site de rencontre), mais l'origine de la photographie. Cohen d'abord (2017), et Agüera y Arcas *et al.* ensuite (2018), soulignent que ce qui importe est de savoir si la photographie a été autopostée ou pas. Les effets des stratégies d'autoprésentation en ligne peuvent en effet être communs aux deux sources.

Plusieurs recherches sur le « gaydar » humain avaient déjà observé des variations dans la performance selon l'origine des photographies (Cox *et al*., 2016 ;

Rule et Ambady, 2008). Si la capacité de discrimination des juges humains confrontés à des photographies prises en conditions contrôlées n'est pas meilleure que le hasard, lorsque les photographies faciales utilisées sont autopostées sur Facebook, en revanche, la capacité à discriminer augmente. Toutefois, lorsque les photographies sont postées par des amis et non pas par l'individu lui-même, la performance diminue à nouveau. Ceci serait particulièrement pertinent pour des gays ouvertement gays. Les images des utilisateurs masculins de Facebook qui ont indiqué avoir un partenaire de même sexe et qui ont aimé au moins deux pages dans un ensemble de pages telles que « Manhunt » ou « J'aime être gay », tout comme les images des utilisateurs du site de rencontres, sont toutes des images autopostées et assumées comme telles. Il n'est ainsi pas surprenant que le modèle ne parvienne pas à les discriminer.

Critique de la transposition de la machine du laboratoire dans l'espace public à l'aune de l'AUC et du problème des faux positifs

Suivant une approche standard d'évaluation des classificateurs binaires, Kosinski et Wang se sont servis de l'AUC. L'article présente néanmoins des ambiguïtés dans la manière de se référer à la performance du modèle sur la base de l'AUC. Quelques commentaires et critiques ont fait remarquer que les performances de leur modèle ne peuvent pas être interprétées comme la probabilité de bien classifier une photo (Bjork-James, 2017 ; Cohen, 2017 ; Howard, 2017). La description de l'« exactitude » (*accuracy*) du modèle en pourcentages sur la base des scores AUC serait ainsi peu appropriée et surtout trompeuse. Le lecteur non averti tombe facilement dans le « piège » : car l'expression « exact à 91 % » traduit mal la signification d'une AUC de 0,91. Les auteurs misent pourtant sur cette ambiguïté, passant sous silence les conditions de l'interprétation probabiliste des scores obtenus, en affirmant dans le résumé que leur modèle peut distinguer correctement une personne gay ou lesbienne d'une personne hétérosexuelle dans respectivement 81 % et 71 % des cas à partir d'une seule image[20], créant ainsi les conditions de l'alerte. Il s'agit pourtant d'une situation hypothétique dans laquelle le modèle devrait choisir entre une image de personne homosexuelle et une image de personne hétérosexuelle. Une telle situation est tout à fait inapplicable dans l'espace

20. « Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women. » Notre traduction : « Avec une seule image faciale, un classificateur peut distinguer correctement les hommes homosexuels des hommes hétérosexuels dans 81 % des cas et dans 71 % des cas chez les femmes. » (Wang et Kosinski, 2018).

public, où il n'y a que des individus, dont la prévalence d'homosexuels est de surcroît relativement faible (de l'ordre de 6 à 7 %).

Kosinski et Wang expliquent pourtant dans la discussion ce qu'il faut comprendre par une telle AUC et, par un jeu rhétorique, ils tentent de retourner la faible prévalence à leur avantage. Ils mobilisent alors l'interprétation plus conventionnelle de l'AUC, dans laquelle un score est attribué à chaque individu, en proposant une situation fictive lors de laquelle ils sélectionnent les 10 % des hommes dont les scores relatifs à l'homosexualité seraient les plus élevés. Nous l'avons souligné : dans un tel cas, la moitié des hommes seraient « réellement homosexuels » (56 %), ce qui signalerait, avec une prévalence de 7 % d'hommes gays, un modèle 8 fois plus performant que le hasard. Kosinski et Wang éludent néanmoins le risque pourtant considérable de faux positifs qui compromettrait totalement le modèle de prédiction visant à identifier un groupe minoritaire (et, inversement, la nullité du modèle dans le cas où la sensibilité, alors trop élevée, produirait essentiellement des faux négatifs).

Aux nombreux doutes quant à la capacité du modèle prédictif de l'orientation sexuelle à s'adapter à un autre environnement que celui sur lequel il a été entraîné, et plus particulièrement à des images qui ne s'inscriraient pas dans une perspective d'autoprésentation, s'ajoute ainsi le risque de ne pas saisir que sa performance serait compromise dans l'espace public étant donné la prévalence relativement faible de l'homosexualité.

CONCLUSION

Pour conclure, il est pertinent de revenir sur l'agitation médiatique qu'avaient déjà générée en octobre 2016 les revendications faites par Faception, start-up israélienne de services en analyse des risques pour des gouvernements et des entreprises pour laquelle Kosinski avait été conseiller en matière d'éthique (Lubin, 2016). Faception affirmait alors pouvoir repérer des sujets dangereux grâce au profilage facial basé sur l'apprentissage profond à partir de photographies de visages de terroristes, de pédophiles, de joueurs de poker professionnels, ainsi que de documents et de dossiers administratifs. Comme la machine prédictive de Kosinski et Wang, celle de Faception attribue des scores de probabilités à partir d'un large éventail de variables, dont la structure du visage et d'autres traits physiques qui résulteraient d'« influences prénatales ». Il serait ainsi possible d'associer l'appartenance d'un nouveau visage à un type donné avec une exactitude de 80 % (Lubin, 2016). La machine de Faception n'était

pas loin de la performance de celle de Kosinski, mais la critique s'est surtout engagée à montrer que c'est plutôt l'inefficacité du système de prédiction basé sur du profilage et l'importance des faux positifs qui étaient inquiétants. Considérant le nombre inacceptable de faux positifs au regard de la prévalence des catégories concernées (terroristes, pédophiles, joueurs de pokers, etc.), la menace ne portait pas tant sur les catégories de populations ciblées que sur *l'ensemble de la population*.

Cette crainte fut d'ailleurs clairement relayée par *Business Insider*, sur la base du témoignage de divers experts du domaine (Lubin, 2016), dont Alexander Todorov, psychologue de Princeton spécialisé dans la perception *humaine* des visages, qui a cosigné avec les membres de l'équipe de Google la première contre-étude visant à mettre à mal une partie des conclusions de Kosinski et Wang (Agüera y Arca *et al.*, 2018). Selon Todorov, même si les performances atteintes par de tels classificateurs sont supérieures au hasard, les faux positifs sont inévitables et inacceptables étant donné la faible prévalence des qualités recherchées (propos recueillis par Lubin, 2016). Or ce furent aussi les propos de Michal Kosinski lui-même, en des termes encore plus explicites. Pour Kosinski, « même le modèle le plus précis visant un résultat rare produira une grande majorité de faux positifs – des "vrais positifs" extrêmement rares, comme être un terroriste, seront cachés parmi des milliers ou des centaines de milliers de "faux positifs" » (propos de Kosinski recueillis par Lubin, 2016, notre traduction). À ceci Kosinski ajoutait les risques d'encoder des biais et de surreprésenter les stéréotypes, mais aussi la difficulté des algorithmes à dissocier l'ethnicité, le sexe et d'autres facteurs, pouvant conduire à des profilages inacceptables. Il est dès lors intéressant de noter que Kosinski renvoie à la lecture de ses propres critiques à propos de Faception, dans l'*authors' note* accompagnant l'étude sur la détection de l'orientation sexuelle. Aussi, comment comprendre que Kosinski se permette en 2017 de généraliser des résultats expérimentaux circonscrits et imprécis alors qu'il dénonçait cette pratique en 2016 ?

Pour y répondre, on peut émettre l'hypothèse suivante. L'alerte qui conclut le papier de Wang et Kosinsky et le traitement médiatique qui s'ensuivit constituent une stratégie rhétorique – presque indispensable – pour assurer la recevabilité d'une telle publication. Afin de ne pas s'exposer à l'accusation de complicité et d'irresponsabilité, les auteurs doivent à la fois énoncer des résultats de leur recherche et alerter sur les dangers qu'ils recouvrent. En produisant cet énoncé sous contrainte, ils peuvent tirer un double bénéfice : disciplinaire, du fait de la performance de leur modèle ; et médiatique, du fait

de l'alerte sur les risques qu'elle poserait pour la vie privée. Cette démarche se traduit concrètement par les montées en généralité abusives de Kosinski et Wang dans le titre, le résumé, l'introduction et la conclusion. C'est pourquoi une analyse des écarts entre le protocole et les énoncés suggère que les auteurs ont préparé la réception de leur recherche, en anticipant la quête de sujets sensibles et provocants des producteurs d'actualité.

Dans cet article, nous avons interrogé la pertinence de l'alerte lancée par Kosinski, en dissociant ses trois composantes (performance du *deep learning* dans la perception, naturalisation de l'orientation sexuelle et transposition à l'espace public). Nous avons pour ce faire systématisé la critique telle qu'elle s'est déployée autour du pre-print de l'article de Kosinski et Wang – qui a finalement été publié dans l'une des revues les plus prestigieuses de psychologie –, en cherchant autant que possible à la structurer, à la rassembler et à l'associer au dispositif méthodologique mobilisé, afin de circonscrire sa portée et plus encore ses limites. Nous avons ainsi questionné le statut de « lanceur d'alerte » que les auteurs revendiquent et mobilisent systématiquement pour justifier la publication de leur recherche, en montrant le décalage entre la performance locale de la recherche menée par Kosinski et Wang et la supposée portée générale de son application, qui justifierait l'alerte.

Si le recours à l'alerte pour justifier la pertinence de recherches contestables n'est pas nouveau, la proposition de Kosinski et Wang est en revanche symptomatique des défis qui se posent aux sciences sociales avec le développement rapide du *machine learning* appliqué aux pratiques sociales. Car si l'alerte formulée par les auteurs se révèle fragile, l'efficience du modèle de prédiction de l'orientation sexuelle proposé n'en demeure pas moins importante, et le débat qui s'ensuivit témoigne de la richesse des enjeux relatifs au déploiement de telles « machines à prédire ». Un premier enjeu tient à la valeur qui peut être accordée à des dispositifs susceptibles d'identifier des structures et des régularités qui échappent à nos modalités de perceptions et d'enquêtes conventionnelles : en d'autres termes, il convient de saisir ce que perçoivent de tels dispositifs malgré la complexité inhérente aux réseaux de neurones profonds, ainsi qu'aux protocoles expérimentaux qui rendent possible leur mise en œuvre dans une perspective de sciences sociales. Un deuxième enjeu porte sur les interprétations de ces régularités, dès lors que nous assistons à la résurgence des présupposés positivistes qui prévalaient aux fondements de la sociologie à la fin du XIX^{e} siècle (Beaude, 2018). En négligeant la réflexivité au profit de déterminations naturelles, c'est plus d'un siècle de sciences sociales qui se trouvent négligées, dans une quête renouvelée de lois sociales,

dès lors que les données seraient abondantes et les capacités de calcul considérablement accrues.

Or la recherche de Kosinski et Wang montre à quel point la puissance inductive rendue possible par le *machine learning* encourage des chercheurs à mobiliser *a posteriori* des théories déterministes pour justifier la performance de leur modèle, sans que le lien ne soit vraiment justifié. En refusant d'expliquer le social par le social, de telles machines prédictives encourent pourtant le risque de surinterpréter leur performance et de généraliser leur capacité au-delà des conditions spécifiques de leur apprentissage. En négligeant non seulement la réflexivité des individus, mais aussi la relative stabilité des normes sociales, le risque est en cela important d'accorder à de telles machines des capacités prédictives qui débordent largement leurs capacités effectives, pourtant aussi circonscrites, instables et provisoires que ne le sont les faits sociaux (Beaude, 2018 ; Jensen 2018).

Au terme de cette analyse, cette machine prédictive se révèle inapte à démontrer les origines hormonales prénatales de l'orientation sexuelle et plus encore à distinguer les orientations sexuelles dans l'espace public. S'il en est une, l'alerte est plutôt celle du risque que des gouvernements et des entreprises utilisent un tel classificateur dans l'espace public alors que la pertinence de son déploiement dans un tel contexte n'est pas démontrée. En mettant l'accent sur la PHT et la vie privée, Kosinski et Wang ont négligé d'autres interprétations possibles de la performance de leur modèle qui ne s'inscrivaient pas dans leur programme de recherche. La capacité des dispositifs algorithmiques à percevoir des normes sociales et non des déterminations naturelles constituait pourtant une piste autrement plus constructive, mais elle aurait posé le débat en des termes probablement incompatibles avec les présupposés de la psychométrie portée par Kosinski.

RÉFÉRENCES

AGÜERA Y ARCAS B., TODOROV A., MITCHELL M. (2018), « Do algorithms reveal sexual orientation or just expose our stereotypes? », *Medium* (blog), 11 janvier.

AMBADY N., HALLAHAN M., CONNER B. (1999), « Accuracy of judgments of sexual orientation from thin slices of behavior », *Journal of Personality and Social Psychology*, vol. 77, n° 3, pp. 538-547.

ANDERSON D. (2017), « GLAAD and HRC call on Stanford University & responsible media to debunk dangerous & flawed report claiming to identify LGBTQ people through facial recognition technology », *GLAAD*, communiqué de presse, 8 septembre 2017.

BAILEY J. M., VASEY P. L., DIAMOND L. M., BREEDLOVE S. M., VILAIN E., EPPRECHT M. (2016), « Sexual Orientation, Controversy, and Science », *Psychological Science in the Public Interest*, vol. 17, n° 2, pp. 45-101.

BEAUDE B. (2018), « (re)Médiations numériques et perturbations des sciences sociales contemporaines », *Sociologie et sociétés*, vol. 49, n° 2.

BEER D. (2017), « The social power of algorithms », *Information, Communication & Society*, vol. 20, n° 1, pp. 1-13.

BENBOUZID B. (2017), « Des crimes et des séismes. La police prédictive entre science, technique et divination », *Réseaux*, n° 206, pp. 95-123.

BERGSTROM C., WEST J. (2017), « Case Study – Machine learning about sexual orientation? », *Calling Bullshit* (site web), 19 septembre 2017.

BJORK-JAMES C. (2017), « Bad science journalism: Gay facial recognition », *Carwil without Borders* (blog), 9 septembre 2017.

BOLLINGER A. (2017), « HRC and GLAAD release a silly statement about the 'gay face' study », *LGBTQ Nation*, 10 septembre 2017.

BURRELL J. (2016), « How the machine 'thinks': Understanding opacity in machine learning algorithms », *Big Data & Society*, vol. 3, n° 1, pp. 1-12.

CARDON D. (2015), *À quoi rêvent les algorithmes : nos vies à l'heure des big data*, Paris, Seuil.

CARDON D. (2018), « Le pouvoir des algorithmes », *Pouvoirs*, n° 164, pp. 63-73.

CASILLI A. (2017), « Une intelligence artificielle révèle les préjugés anti-LGBT des chercheurs de Stanford », *Antonio A. Casilli* (blog), 9 septembre 2017.

CHATEAURAYNAUD F. (2013), « Lanceur d'alerte », in I. CASILLO, R. BARBIER, L. BLONDIAUX, F. CHATEAURAYNAUD, J.-M. FOURNIAU, R. LEFEBVRE, C. NEVEU et D. SALLES (dir.), *Dictionnaire critique et interdisciplinaire de la participation*, Paris, GIS Démocratie et Participation.

CHATEAURAYNAUD F., TORNY D. (1999), *Les sombres précurseurs : une sociologie pragmatique de l'alerte et du risque*, Paris, Éditions de l'EHESS.

COHEN P. N. (2017), « On artificially intelligent gaydar », *Family Inequality* (blog), 11 septembre 2017.

COX W. T. L., DEVINE P. G., BISCHMANN A. A., HYDE J. S. (2016), « Inferences About Sexual Orientation: The Roles of Stereotypes, Faces, and The Gaydar Myth », *The Journal of Sex Research*, vol. 53, n° 2, pp. 157-171.

DIAKOPOULOS N. (2014), « Algorithmic-Accountability: the investigation of Black Boxes », *Tow Center for Digital Journalism*.

FACEBOOK INC. (2018), « Responses To Judiciary Committee Questions For The Record », document du 8 juin 2018 soumis dans le cadre de l'audience titrée : « Facebook, Social Media Privacy, and the Use and Abuse of Data », devant l'United States Senate Committee on the Judiciary, tenue le 10 avril 2018.

FARA P. (2008), « Marginalized Practices », in Roy PORTER (ed.), *The Cambridge History of Science*, Cambridge, Cambridge University Press, pp. 485-506.

FLAHERTY C. (2017), « AI Gaydar Study Gets Another Look », *Inside Higher Ed*, 13 septembre 2017.

GELMAN A., MATTSON G., SIMPSON D. (2018), « Gaydar and the Fallacy of Decontextualized Measurement », *Sociological Science*, vol. 5, pp. 270-280.

GERSHGORN D. (2017), « A Stanford scientist says he built a gaydar using "the lamest" AI to prove a point », *Quartz*, 16 septembre 2017.

GRASSEGGER H., KROGERUS M. (2017), « The Data That Turned the World Upside Down », *Motherboard*, 28 janvier 2017.

HAWKINS D. (2017), « Researchers use facial recognition tools to predict sexual orientation. LGBT groups aren't happy », *Washington Post*, 12 septembre 2017.

HOFMAN J. M., SHARMA A., WATTS D. J. (2017), « Prediction and explanation in social systems », *Science*, vol. 355, n° 6324, pp. 486-488.

HOWARD J. (2017), « Can Neural Nets Detect Sexual Orientation? A Data Scientist's Perspective », *Fast.ai* (site web), 13 septembre 2017.

INTRONA L., NISSENBAUM H. (2010), « Facial Recognition Technology: A Survey of Policy and Implementation Issues », *Organisation, Work and Technology Working Paper Series*, Lancaster, Lancaster University, The Department of Organisation, Work and Technology.

JANNINI E. A., BLANCHARD R., CAMPERIO-CIANI A., BANCROFT J. (2010), « Male homosexuality: nature or culture? », *The Journal of Sexual Medicine*, vol. 7, n° 10, pp. 3245-3253.

JENSEN P. (2018), *Pourquoi la société ne se laisse pas mettre en équations*, Paris, Seuil.

JORDAN-YOUNG R. M. (2011), *Brain Storm: The Flaws in the Science of Sex Differences*, reprint edition, Cambridge MA, Harvard University Press.

KLAUSER F. (2016), *Surveillance and Space*, London, Sage.

KOSINSKI M., BEHREND T. (2017), « Editorial overview: Big data in the behavioral sciences », *Current Opinion in Behavioral Sciences*, vol. 18, pp. iv-vi.

KOSINSKI M., MATZ S. C., GOSLING S. D., POPOV V., STILLWELL D. (2015), « Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines », *The American Psychologist*, vol. 70, n° 6, pp. 543-556.

KOSINSKI M., STILLWELL D., GRAEPEL T. (2013), « Private traits and attributes are predictable from digital records of human behavior », *Proceedings of the National Academy of Sciences*, vol. 110, n° 15, pp. 5802-5805.

KOSINSKI M., WANG Y. (2017a), « Response to GLAAD and HRC », *Google Docs* (document en ligne), 9 septembre 2017.

KOSINSKI M., WANG Y. (2017b), « Authors' note: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images », *Google Docs*, version du 28 septembre 2017.

KOSINSKI M., WANG Y., LAKKARAJU H., LESKOVEC J. (2016), « Mining big data to extract patterns and predict real-life outcomes », *Psychological Methods*, vol. 21, n° 4, pp. 493-506.

KUANG C. (2017), « Can A.I. Be Taught to Explain Itself? », *The New York Times*, 21 novembre 2017.

LE BRETON D. (1992), *Des visages. Essai d'anthropologie*, Paris, Métailié.

LEVIN S. (2017a), « New AI can work out whether you're gay or straight from a photograph », *The Guardian*, 7 septembre 2017.

LEVIN S. (2017b), « LGBT groups denounce "dangerous" AI that uses your face to guess sexuality », *The Guardian*, 9 septembre 2017.

LEWIS P. (2018), « "I Was Shocked It Was so Easy": meet the Professor Who Says Facial Recognition can Tell If You're Gay ». *The Guardian*, 7 juillet 2018.

LIBÉRATION (2017), « Quand une intelligence artificielle est instrumentalisée pour cibler et essentialiser les gays », *Libération.fr*, 11 septembre 2017.

LUBIN G. (2016), « "Facial-profiling" could be dangerously inaccurate and biased, experts warn », *Business Insider*, 12 octobre 2016.

MATTSON G. (2017), « Artificial Intelligence Discovers Gayface. Sigh », *Greggor Mattson* (blog), 9 septembre 2017.

MILLER A. E. (2018), « Searching for gaydar: Blind spots in the study of sexual orientation perception », *Psychology & Sexuality*, vol. 9, n° 3, pp.188-203.

MURPHY H. (2017), « Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine », *The New York Times*, 9 octobre 2017.

MUSIANI F. (2015), « Edward Snowden, l’“homme-controverse” de la vie privée sur les réseaux », *Hermès, La Revue*, n° 73, pp. 209-215.

NORVAL A., PRASOPOULOU E. (2017), « Public faces? A critical exploration of the diffusion of face recognition technologies in online social networks », *New Media & Society*, vol. 19, n° 4, pp. 637-654.

PARKHI O. M., VEDALDI A., ZISSERMAN A. (2015), « Deep Face Recognition », *BMVC*, vol. 1, pp. 6.

RULE N. O., AMBADY N. (2008), « Brief exposures: Male sexual orientation is accurately perceived at 50ms », *Journal of Experimental Social Psychology*, vol. 44, n° 4, pp. 1100-1105.

RULE N. O., MACRAE C. N., AMBADY N. (2009), « Ambiguous Group Membership Is Extracted Automatically From Faces », *Psychological Science*, vol. 20, n° 4, pp. 441-443.

SKORSKA M. N., GENIOLE S. N., VRYSEN B. M., MCCORMICK C. M., BOGAERT A. F. (2015), « Facial Structure Predicts Sexual Orientation in Both Men and Women », *Archives of Sexual Behavior*, vol. 44, n° 5, pp. 1377-1394.

THE ECONOMIST (2017a), « What machines can tell from your face: Nowhere to hide », *The Economist*, 9 septembre 2017.

THE ECONOMIST (2017b), « Advances in AI are used to spot signs of sexuality: Facial technology », *The Economist*, 9 septembre 2017.

VINCENT J. (2017), « The invention of AI ‘gaydar’ could be the start of something much worse », *The Verge*, 21 septembre 2017.

WANG Y., KOSINSKI M. (2018), « Deep neural networks are more accurate than humans at detecting sexual orientation from facial images », *Journal of Personality and Social Psychology*, vol. 114, n° 2, pp. 246-257.

WEBER C. (2017), « The Face of Sexuality: Why Do AI-Generated Sexual Orientations Matter? », *The Disorder of Things* (blog), 25 septembre 2017.

ZIEWITZ M. (2016), « Governing Algorithms Myth, Mess, and Methods », *Science, Technology & Human Values*, vol. 41, n° 1, pp. 3-16.