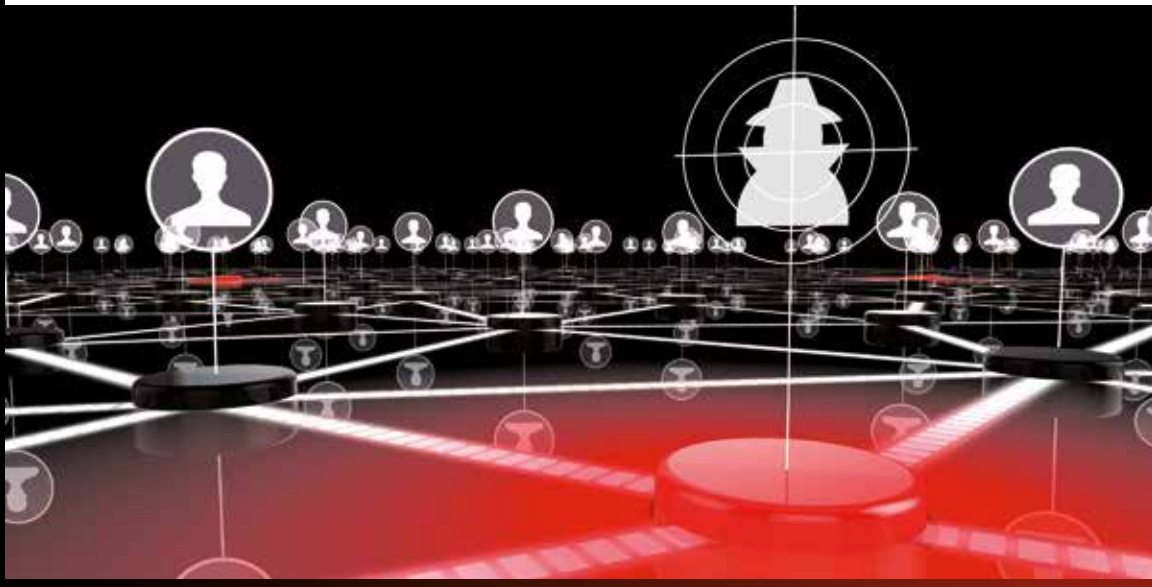
Sous la direction de
SCHALLUM PIERRE
FEHMI JAAFAR

MÉDIAS SOCIAUX

PERSPECTIVES SUR LES DÉFIS LIÉS À LA CYBERSÉCURITÉ, LA GOUVERNEMENTALITÉ ALGORITHMIQUE ET L'INTELLIGENCE ARTIFICIELLE



ÉTHIQUE IA ET SOCIÉTÉ Collection

PUL

ÉTHIQUE IA ET SOCIÉTÉ

Collection dirigée par Lyse Langlois

Médias sociaux

Perspectives sur les défis liés à la cybersécurité,
la gouvernementalité algorithmique
et l'intelligence artificielle

Médias sociaux

Perspectives sur les défis liés à la cybersécurité,
la gouvernementalité algorithmique
et l'intelligence artificielle

Sous la direction de
Schallum Pierre et Fehmi Jaafar

Presses de
l'Université Laval

Financé par le gouvernement du Canada
Funded by the Government of Canada

Canada

Nous remercions le Conseil des arts du Canada de son soutien.
We acknowledge the support of the Canada Council for the Arts.

Financé par le gouvernement du Canada
Funded by the Government of Canada

Canada

Les Presses de l'Université Laval reçoivent chaque année de la Société de développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

SODEC
Québec

Maquette de couverture : Laurie Patry
Mise en page : In Situ

Dépôt légal 4e trimestre 2020

ISBN 978-2-7637-5328-7
PDF 9782763753294

Les Presses de l'Université Laval
www.pulaval.com

TABLE DES MATIÈRES

Introduction 1
SCHALLUM PIERRE ET FEHMI JAAFAR

1. Social Network Cyberattacks and Security Issues 11
FEHMI JAAFAR AND JEAN DECIAN

2. La Sécurité utilisable : Entre l'interaction homme-machine et la sécurité de l'information 23
HERVÉ SAINT-LOUIS

3. Protéger autant nos données que les fondements de la démocratie en privilégiant les logiciels libres 45
MATHIEU GAUTHIER-PILOTE

4. Access Control in Cybersecurity and Social Media 69

5. Social Media Surveillance : Between Digital Governmentality, Big Data and Computational Social Science 107
RAMÓN REICHERT

6. Technique, société et cyberespace : La gouvernementalité algorithmique 121
MARC MÉNARD ET ANDRÉ MONDOUX

7. The Case of Fake News and Automatic Content Generation in the Era of Big Data and Machine Learning 139
NICOLAS GARNEAU

8. Protecting Online Communities from Harmful Behaviors 149
MARC-ANDRÉ LAROCHELLE, ÉLOI BRASSARD-GOURDEAU, ZEINEB TRABELSI, RICHARD KHOURY, SEHL MELLOULI, LIZA WOOD, CHRIS PRIEBE

9. Extrémisme violent de droite et médias sociaux : caractéristiques, idéologies, médiatisations et gestions 167
SCHALLUM PIERRE

La publication du présent livre a été rendue possible grâce à un financement de l'Institut intelligence et données (IID) et du Centre de recherche informatique de Montréal (CRIM).

INTRODUCTION

Schallum Pierre et Fehmi Jaafar

Schallum Pierre est chargé scientifique et éthique à l'Institut intelligence et données (IID) de l'Université Laval et professeur à temps partiel de "communications sociales et médias sociaux " à l'Université Saint-Paul. Chercheur en éthique des données massives, il s'intéresse à la question de l'identité dans ses dimensions technologiques, numériques, anthropologiques, idéologiques et historiques. Il est membre du comité de rédaction de la revue *Technologie et innovation*, membre du Réseau intégré sur la cybersécurité de l'Université de Montréal et membre du comité d'éthique du CHU de Québec-Université Laval. Détenteur d'un doctorat en philosophie de l'Université Laval, il a effectué un stage postdoctoral à Polytechnique Montréal.

Fehmi Jaafar est chercheur au Centre de recherche en Informatique de Montréal (CRIM) et professeur adjoint affilié à l'Université Concordia. Il est le Vice-président du comité sur l'Internet des objets et technologies connexes au Conseil canadien des normes. Auparavant, il a été professeur adjoint à l'Université Concordia d'Edmonton, et chercheur postdoctoral à Queen's University et à Polytechnique Montréal. Après avoir obtenu un doctorat en informatique de l'Université de Montréal, M. Jaafar s'est spécialisé dans des travaux de recherche en cybersécurité, Il s'intéresse à la cybersécurité dans l'Internet des objets et à l'application des techniques d'apprentissage automatique en cybersécurité. Il a établi des programmes de recherche en collaboration avec Défense Canada, Sécurité publique Canada, le Conseil de recherches en sciences naturelles et en génie du Canada, et des partenaires industriels et universitaires canadiens et étrangers.

La cybersécurité dans le contexte des médias sociaux est souvent rattachée à la gestion des risques de l'information. En ce sens, elle réfère à la norme ISO/CEI 27001 :2013(fr) qui préconise la préservation de la confidentialité, de l'intégrité et de la disponibilité de l'information. Le système de management de la sécurité de tout média social est de première importance car il y va de la confiance qui lui sera accordée. Une perte de confiance peut engendrer une baisse de l'utilisation ou de l'adoption d'une plateforme. Elle pourrait même mener à son déclin. Les enjeux de sécurité sont susceptibles d'atteindre non seulement la fréquentation des médias sociaux, mais aussi et surtout, les données personnelles. À ce titre, l'affaire Cambridge Analytica est riche d'enseignement. Les données de 87 millions de personnes ont été collectées, sans l'obtention de leur consentement libre et éclairé, à des fins de désinformation et d'influence idéologique (Isaak et Hanna, 2018). Facebook a été utilisé comme outil de manipulation de l'opinion publique et de propagation à outrance de rumeurs. L'affaire montre les conséquences sur le plan politique d'une exploitation massive des données, portant ainsi une grave atteinte tant à la vie privée qu'à un régime dit démocratique (Barraud, 2018).

La cybersécurité peut être causée par des défaillances de procédures organisationnelles et des défaillances technologiques. À propos de Facebook, le site dédié à la cybersécurité, Krebsonsecurity.com révèle les conséquences d'une défaillance de procédures organisationnelles sur des millions d'utilisateurs et d'utilisatrices dont les mots de passe ont été stockés en clair pendant plusieurs années à l'interne (Krebsonsecurity, 2019). En plus des défaillances de procédures organisationnelles, il faut noter des défaillances technologiques. Selon les statistiques disponibles sur le site Downdetector.com, les défaillances technologiques sont un problème récurrent auquel les médias sociaux sont confrontés. À la date du 20 juillet 2020, par exemple, les principaux problèmes signalés par les utilisateurs et utilisatrices de Facebook concernent la connexion (38 %), les images (30 %) et les interruptions totales (30 %) (Downdetector, 2020). Ces problèmes sont à l'origine de fréquentes indisponibilités. D'autres grandes plateformes ont récemment connu des défaillances technologiques. La panne majeure qu'ont expérimenté les utilisateurs et utilisatrices des services de Google le 2 juin 2019 l'illustre bien. Des plateformes comme Discord, Snapchat ou Youtube utilisant les services de Google Cloud Platform ont éprouvé des problèmes de connexion, de congestion et de performance (Lefaix, 2019). Pendant plus de 4 heures, de nombreuses applications ont été indisponibles pour des millions d'utilisateurs et

d'utilisatrices. On pourrait aisément imaginer les conséquences sur la vie humaine si ces plateformes, se basant sur des systèmes centralisés, devaient être les principaux moyens de paiement pour l'accès en temps réel aux soins de santé.

Mais, que sont au juste les médias sociaux ? Ce sont des technologies et des services qui rendent possible le réseautage entre individus. Se caractérisant par l'interaction et le partage entre les pairs, ils constituent un moment important de l'évolution du Web, souvent appelés Web 2.0. Dans cette perspective, contrairement au Web 1.0 qui est statique et unidirectionnel, le Web 2.0, dont les médias sociaux en sont la manifestation, est dynamique et bidirectionnel. Autrement dit, alors que le rapport au Web 1.0 se limite soit à un statut de producteurs/productrices ou de consommateurs/consommatrices de contenu, le Web 2.0 se fonde sur la collaboration des membres d'une communauté (Nath, Dhar et Basishtha, 2014). Les concepts comme la participation, l'effet de réseau, l'horizontalité, l'intelligence collective, sont quelques-uns des attributs des médias sociaux. Dans le Web 2.0, il n'y a pas, comme pour le Web 1.0, d'un côté les auteurs/autrices et de l'autre les lecteurs/lectrices car le message diffusé peut être coproduit, échangé et modifié par une communauté.

Les médias sociaux ont transformé de nombreuses professions dont la pratique du journalisme. Jusqu'à la fin des années 2000, par exemple, les journaux pouvaient se limiter à leur version papier pour satisfaire leur lectorat. À partir des années 2010, ils doivent compter sur les médias sociaux pour attirer de nouveaux lecteurs et de nouvelles lectrices, devenus, à leur tour, des rédacteurs et des rédactrices. C'est que les médias sociaux, par le temps réel et l'ouverture à la participation, ont changé totalement le rapport des journaux avec leur audience. Les abonnés et les abonnées sont libres « d'éditer et de publier des textes, des liens, des images, des photographies, des vidéos et des enregistrements sonores », comme c'est le cas par exemple avec le club de Mediapart qui est un « blog d'information participatif » (Mediapart, 2019).

L'usage des médias sociaux, qui connaît un bond sans précédent, prend une place de plus en plus prépondérante dans la vie en ligne. Partout à travers le monde, les jeunes comme les adultes, utilisent les plateformes de socialisation. Au Canada, pas moins de 75 % des adultes restent quotidiennement 3 à 4 heures en ligne (ACEI, 2019). Au Québec, l'usage des médias sociaux est maintenant de 98 % chez les 18 à 24 ans (CEFRIO, 2018). Les plateformes de réseautage social en ligne ont un intérêt pratique.

Elles sont utilisées pour initier ou préserver des liens amicaux, familiaux, amoureux et collégiaux. Elles mobilisent l'espace, le corps et le lien social définissant les « sociabilités numériques » (Casilli, 2010 : 12).

Les médias sociaux sont utilisés par certains pays pour démontrer leur capacité à produire ou à réagir à des cyberattaques, lesquelles sont des agressions visant à endommager, de façon durable ou non, un réseau. Un pays dont les médias sociaux sont victimes de cyberattaques peut révéler ses limites en termes de compétences technologiques. Les vulnérabilités témoignent d'un grave problème de contrôle de son cyberespace voire de sa souveraineté numérique. Il y a un jeu idéologique de puissance qui s'observe et s'impose dans l'orientation et le développement des médias sociaux. Avec ses 2,38 milliards d'utilisateurs et d'utilisatrices, Facebook (Clément, 2019) domine les médias sociaux. Cette situation est caractéristique d'une idéologie de domination de la Silicon Valley (Alloa et Soufron, 2019). À cet effet, l'affaire Cambridge Analytica pourrait être considérée comme une attaque de l'hégémonie de Facebook.

Dans ce contexte, les cyberattaques renvoient à la cyberdéfense et à la géopolitique. Impliquant des stratégies offensives et défensives, les médias sociaux constituent un terreau de guerre ou « un champ de bataille » comme le rappelle très justement le lanceur d'alerte Christopher Wylie dont les révélations ont été à l'origine du scandale Facebook-Cambridge Analytica (Wylie, 2020 : 422). Ils définissent le lieu de nouveaux enjeux, confirmant cette pensée du célèbre stratège militaire Carl von Clausewitz « War is a mere continuation of policy by other means ». (Clausewitz, 1943 : 16) La Chine, la Russie et la Corée du nord sont les puissances montantes de cette cyberguerre utilisant l'intelligence artificielle. WannaCry, Adylkuzz, Petya, NotPetya sont des noms de cyberattaques ayant eu de graves conséquences sur des infrastructures technologiques, à l'échelle nationale ou mondiale. Avec les nouveaux outils performants de piratage, le cyberespace est devenu un haut lieu de cyber-conflictualité sans précédent (Nocetti, 2018).

Le pouvoir de la propagation rapide des informations sur les médias sociaux peut être utilisé comme moyen de pression, voire comme force de frappe. Sur le plan géopolitique, l'usage de Twitter peut avoir des impacts considérables sur les relations diplomatiques. Les remous qu'a eu un tweet de Chrystia Freeland, la ministre des Affaires étrangères du Canada, sur la diplomatie entre l'Arabie saoudite et le Canada en est un exemple. Le 2 août 2018, sur son compte certifié, madame Freeland avait appelé à la libération de Raif et Samar Badawi (Freeland, 2018). Les dispositions prises par

l'Arabie Saoudite, comme le rappel de son ambassadeur et le gel d'échanges commerciaux, sont là pour montrer à quel point les médias sociaux peuvent être un outil stratégique ou une arme entre les mains des états. La rapidité avec laquelle ces informations ont été diffusées et partagées pose le problème de l'utilisation des données massives. Les données massives sont caractérisées par la règle des 5 V que sont la vélocité ou la vitesse à laquelle les données sont générées, la variété ou les différentes sources des données, la véracité ou la qualité de la donnée et la valeur ou l'interprétation de la donnée (Gupta, Kumar, and Dwivedi, 2018 ; Dautov et Distefano, 2017 ; Sheth, 2014). La gestion, l'analyse et le calcul prédictif de données massives peuvent servir à orienter l'opinion et les prises de position des utilisateurs et utilisatrices.

La sécurité devient essentielle si les médias sociaux sont le lieu où les données à caractère personnel sont utilisées et échangées. Comment s'assurer que le média social dont on se sert ou auquel on contribue respecte et protège la vie privée des individus ? Quelle politique appliquer pour authentifier les profils ? Jusqu'à quel point peut-on croire une information publiée sur un compte personnel ?

La cybersécurité concerne non seulement le respect des données à caractère personnel mais aussi les données liées à la vie privée. En effet, l'un des enjeux des médias sociaux est l'exploitation des données qui peut conduire à l'extraction de renseignements d'ordre privé. Les données sont au cœur de cette course au biopouvoir, lequel renvoie à un contrôle ou une « organisation du pouvoir sur la vie » des individus et des populations (Foucault, 1976 : 183). Le contrôle de la vie ou le pouvoir sur le corps des individus est exercé par des plateformes, dirigées certes par quelques individus mais pouvant être largement favorisés par des états. De nombreux pays occidentaux et gouvernements élus démocratiquement y imposent arbitrairement leurs politiques. Ils exploitent les données massives pour augmenter, de façon ciblée, la surveillance et la censure (Feldstein, 2019).

Les plateformes de réseautage social en ligne ont ceci de particulier, elles contiennent de grandes quantités de données qui sont au cœur de leurs modèles d'affaires. À ce sujet, les données collectées sur les médias sociaux sont utilisées et exploitées pour la publicité ciblée et personnalisée. Tant les entreprises que les gouvernements s'en servent pour influencer ou renforcer des choix qui peuvent être d'ordre économique. L'utilisation des données à caractère personnel pose des questions hautement éthiques, lesquelles ont fait l'objet de la journée d'étude « Cybersécurité et médias

sociaux », tenue à l'Université Saint-Paul d'Ottawa le 2 novembre 2018. Cet ouvrage, qui reprend trois des présentations de cette journée d'étude, propose un regard élargi sur les médias sociaux afin d'approfondir les questions divergentes qui gravitent autour des problématiques de la cybersécurité, de l'intelligence artificielle, de l'éthique, de la sécurité des données privées et des logiciels libres.

Chaque chapitre approfondit une de ces dimensions dans le but de mettre de l'avant les enjeux complexes que suscitent ces développements technologiques. Nous avons permis aux auteurs et auteures d'écrire dans la langue de leur choix.

Le livre se divise en trois parties : dans la première partie, allant du chapitre 1 au chapitre 4, il interroge la définition et la sécurité des médias sociaux. Dans la deuxième partie, allant du chapitre 5 au chapitre 6, le livre traite de la problématique de la gouvernementalité algorithmique et de la surveillance. Enfin, du chapitre 7 au chapitre 9, sont analysés les phénomènes de fausses rumeurs, de comportements nuisibles et de discours haineux.

Le chapitre 1 « Social Network Cyberattacks and Security Issues » dresse le portrait des principales cyberattaques sur les médias sociaux et les problèmes de sécurité signalés depuis 2016. Fehmi Jaafar et Jean Decian analysent, avec minutie, les scénarios et les résultats de ces cyberattaques et problèmes de sécurité afin de déterminer un ensemble de leçons apprises et de recommandations pour améliorer la posture de sécurité des réseaux sociaux.

Dans le chapitre 2 « La Sécurité utilisable : Entre l'interaction homme-machine et la sécurité de l'information », Hervé Saint-Louis souligne l'importance de l'humain. Il explore la place de la vie privée du point de vue de l'interaction homme-machine, du développement des médias sociaux et de la marchandisation des données des utilisateurs et utilisatrices.

Le chapitre 3 « Protéger autant nos données que les fondements de la démocratie en privilégiant les logiciels libres » montre les avantages du logiciel libre pour la sécurité informatique et la protection des utilisateurs et utilisatrices d'appareils numériques. Dans la lignée de Richard Stallman, Mathieu Gauthier-Pilote soutient que les logiciels libres sont essentiels à la préservation des libertés qui sont au fondement de la société démocratique. L'auteur considère le problème de la centralisation d'Internet

et l'impact des nouvelles plateformes décentralisées ou distribuées, devant permettre à mieux contrôler soi-même ses données.

Dans le chapitre 4 « Access Control in Cybersecurity and Social Media », Nadine Kashmara, Mehdi Addaa, Mirna Atiehb et Hussein Ibrahim présentent l'importance des modèles de contrôle d'accès pour la cybersécurité et les médias sociaux, tout en expliquant les services de médias sociaux et les technologies utilisées. Ils analysent les différents types d'attaques et les stratégies pour contrôler l'accès à ces médias.

Le chapitre 5 aborde la question de la gouvernementalité. Dans « Social Media Surveillance : Between Digital Governmentality, Big Data and Computational Social Science », Ramón Reichert examine la façon dont les médias sociaux facilitent la gouvernance algorithmique et l'exercice d'une biosurveillance avec la biométrie.

Dans le chapitre 6 « Technique, société et cyberespace : La gouvernementalité algorithmique » Marc Ménard et André Mondoux reviennent sur le débat de la neutralité de la technologie. Ils mettent en lumière la dimension politique de cette technologie présentée comme neutre. Il se construit un rapport de pouvoir qui détermine les modes de savoirs et d'êtres dont les conséquences sur la sécurité du cyberespace est préoccupante. L'exploitation des données, de façon non transparente, dans des perspectives économiques remettent en question cette neutralité et constituent un facteur aggravant de la sécurité des médias sociaux.

L'intelligence artificielle est aujourd'hui utilisée pour générer automatiquement de « fausses nouvelles ».

Dans le chapitre 7, « The Case of Fake News and Automatic Content Generation in the Era of Big Data and Machine Learning », Nicolas Garneau examine les modèles d'apprentissage machine pour la génération de faux contenus et les recherches préliminaires sur les stratégies de lutte contre ce phénomène émergent.

Le chapitre 8 « Protecting Online Communities from Harmful Behaviors » explore le sujet de la nuisance en ligne et ses défis. Marc-André Larochelle, Éloi Brassard-Gourdeau, Zeineb Trabelsi, Richard Khoury, Sehl Mellouli, Liza Wood et Chris Priebe discutent de sa définition, des ensembles de données, de la détection et des stratégies de protection de la communauté.

Enfin, le livre examine le problème de la propagation des discours haineux et de l'extrémisme violent de droite en ligne. À partir des cas comme les attentats de Christchurch et l'attentat de la grande mosquée de Québec, Schallum Pierre considère, dans le chapitre 9 « Extrémisme violent de droite et médias sociaux : caractéristiques, idéologies, médiatisations et gestions », la façon dont les médias sociaux sont utilisés pour promouvoir la violence en référence à des idéologies extrémistes de droite. Il souligne, d'une part, les enjeux technologiques et éthiques soulevés par la gestion de l'extrémisme violent de droite, avec l'intelligence artificielle et la chaîne de blocs, et, d'autre part, l'importance de miser sur une approche préventive.

Ces différentes contributions témoignent du caractère interdisciplinaire de la recherche sur les médias sociaux, du point de vue de la cybersécurité. S'adressant autant aux personnes expertes en données massives qu'aux institutions privées et publiques qui rédigent des politiques de gestion de données personnelles, l'ouvrage se veut aussi un guide pour sensibiliser le milieu citoyen sur les cyberattaques se rapportant aux médias sociaux et sur les enjeux liés à la protection des données à caractère personnel. Nous espérons que les chapitres qui suivent sauront, par leurs propositions, contribuer à faire des médias sociaux un outil plus respectueux des données citoyennes.

RÉFÉRENCES

ACEI, (2019). *Dossier documentaire sur internet au Canada*. https ://cira.ca/fr/resources/corporation/dossier-documentaire/canadas-internet-factbook-2019, consulté le 7 août 2019.

Alloa, E. et Soufron, J.-B. (2019). « L'idéologie de la Silicon Valley », *Esprit*. https ://esprit.presse.fr/article/emmanuel-alloa-et-jean-baptiste-soufron/l-ideologie-de-la-silicon-valley-42081, consulté le 5 juillet 2019.

Barraud, B. (2018). « Se souvenir de Cambridge Analytica », *La revue européenne des médias et du numérique*, n°48, https ://la-rem.eu/2018/12/se-souvenir-de-cambridge-analytica/, consulté le 7 juillet 2019.

Canahuati, P. (2019). "Keeping Passwords Secure", *Facebook Newsroom*. https ://newsroom.fb.com/news/2019/03/keeping-passwords-secure/, consulté le 3 août 2019.

Casilli, A. (2010). « Les Liaisons numériques : Vers une nouvelle sociabilité ? », Paris, Seuil.

CEFRIO, (2018). « L'usage des médias sociaux au Québec », *NETendances 2018*, vol. 9 nº 9 5. https ://cefrio.qc.ca/media/2023/netendances-2018_medias-sociaux.pdf, consulté le 7 août 2019.

Centre de la sécurité des télécommunications, (2014). « La cybersécurité et les médias sociaux ». https : //www.cse-cst.gc.ca/fr/interactive-media-medias-interactifs/medias-sociaux-cybersecurite, consulté le 7 août 2019.

Clausewitz, von, C. (1943). "On War, translated by O. J. Matthijs Jolles, foreword by Colonel Joseph Greene and preface by Richard McKeon", New York, The Modern Library.

Clément, J. (2019). *Number of Monthly Active Facebook Users Worldwide as of 1st quarter 2019 (in millions)*, last edited June 5, 2019. https : //www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/, consulté le 16 juillet 2019.

Dautov, R. and Distefano, S. (2017). "Quantifying Volume, Velocity, and Variety to Support (Big) Ddata-Intensive Application Development", *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, pp. 2843-2852. doi : 10.1109/BigData.2017.8258252

Downdetector, (2020). *Facebook*. https : //downdetector.ca/status/facebook/, consulté le 20 juillet 2020.

Feldstein, S. (2019). "How Artificial Intelligence Systems Could Threaten Democracy", *The Conversation*. http : //theconversation.com/how-artificial-intelligence-systems-could-threaten-democracy-109698, consulté le 7 août 2019.

Foucault, M. (1976). « Histoire de la sexualité I : la volonté de savoir », Paris, Gallimard.

Freeland, C. (2018). « Compte certifié de Chrystia Freeland », *Twitter*. https : //twitter.com/cafreeland/status/1025030172624515072 ?lang=fr, consulté le 7 août 2019.

Google Cloud Platform, (2019). *Incident #19009*, Google Cloud Networking. https : //status.cloud.google.com/incident/cloud-networking/19009 ?fbclid=IwAR23Gf5pLLdnMVZPeaVBArB5F7FfSyx7JajmrSfUv2BI1oHJCSe0cFY2EcU, consulté le 5 juillet 2019.

Gupta, B., Kumar, A. and Dwivedi, R. K. (2018). "Big Data and Its Applications – A Review", *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida (UP), India, pp. 146-149, doi : 10.1109/ICACCCN.2018.8748743.

Isaak, J. and Hanna, M. J. (2018). "User Data Privacy : Facebook, Cambridge Analytica, and Privacy Protection," in Computer, vol. 51, nº 8, pp. 56-59, doi : 10.1109/MC.2018.3191268

ISO/CEI, (2014). « ISO/CEI 27001 :2013(fr)Technologies de l'information — Techniques de sécurité — Systèmes de management de la sécurité de l'information — Exigences », https : //www.iso.org/obp/ui/fr/#iso :std :iso-iec :27001 :ed-2 :v1 :fr, consulté le 5 juillet 2019.

Krebsonsecurity (2019). *Facebook Stored Hundreds of Millions of User Passwords in Plain Text for Years*, https : //krebsonsecurity.com/2019/03/facebook-stored-hundreds-of-millions-of-user-passwords-in-plain-text-for-years/, consulté le 3 août 2019.

Lefaix, É. (2019). « Une panne de Google Cloud rend indisponible Discord, Snapchat ou encore YouTube », *SiecleDigital*, Publié le 3 juin 2019 à 08h30 - Mis à jour le 3 juin 2019 à 09h29. https : //siecledigital.fr/2019/06/03/panne-google-cloud-impact-snapchat-youtube/, consulté le 17 juillet 2020.

Mediapart, (2019). « Qu'est-ce que le Club ? » https : //www.mediapart.fr/le-club, consulté le 3 juillet 2019.

Nath, K., Dhar, S. and Basishtha, S. (2014). "Web 1.0 to Web 3.0 - Evolution of the Web and its Various Challenges", *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, Faridabad, pp. 86-89, doi : 10.1109/ICROIT.2014.6798297

Nocetti, J. (2018). « Géopolitique de la cyber-conflictualité », *Politique étrangère* 2018/2 (Été), https://www.cairn.info/article.php?ID_ARTICLE=PE_182_0015#xd_co_f=MWY5N2U0ZmItNTY1OC00YTQzLTg4MWItNTkyNDEwODMyNDRl~, consulté le 7 juillet 2019.

Oehri C. and Teufel, S. (2012). "Social Media Security Culture", *2012 Information Security for South Africa*, Johannesburg, Gauteng, pp. 1-5. doi : 10.1109/ISSA.2012.6320436

Sheth, A. (2014). "Transforming Big Data into Smart Data : Deriving Value via Harnessing Volume, Variety, and Velocity Using Semantic Techniques and Technologies," *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, pp. 2-2, doi : 10.1109/ICDE.2014.6816634.

Thuraisingham, B., Kantarcioglu, M. and Khan, L. (2018). "Integrating Cyber Security and Data Science for Social Media : A Position Paper," *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, Vancouver, BC, pp. 1163-1165, doi : 10.1109/IPDPSW.2018.00178

Wylie, C. (2020). *Mindfuck : le complot cambridge analytica pour s'emparer de nos cerveaux*, Paris, Grasset.

1

SOCIAL NETWORK CYBERATTACKS AND SECURITY ISSUES

Fehmi Jaafar and Jean Decian

Fehmi Jaafar received his PhD from the Department of Computer Science at the University of Montreal, Canada. He is specialized in cybersecurity research, notably at Queen's University and Polytechnique Montréeal. Dr. Jaafar is interested in cybersecurity in the Internet of Things, in the evolution, security and quality of software, and in the application of machine learning techniques in cybersecurity. He is currently an adjunct professor at the Faculty of Management at Concordia University of Edmonton and a Researcher at the Computer Research Institute of Montréal (CRIM).

Jean Decian is currently a computer engineering student at Polytechnique Montréal. He is interested in the analysis of cybersecurity attacks related to social media and the Internet of Things. Jean has several years of experience in software development and data analysis. He is a scientific writer willed to communicate scientific information and facts about cyber threats and cyber defense.

ABSTRACT

In the last five years, multiple cyberattacks targeting social networks were reported. Moreover, medias and security experts disseminated a set of serious security issues in social networks. In fact, the amount of personal information and sensitive data that may be managed by social networks make them an ideal target for hackers and hacking organizations. In this chapter, we will review the main social network cyberattacks and security issues reported since 2016. Our goal is to present a deep and accurate analysis of scenarios and results of these cyberattacks and security issues in order to determine a set of lessons learned and recommendations to enhance the security posture of the social networks.

1. INTRODUCTION

In the last five years, multiple cyberattacks targeting social networks were reported. Moreover, medias and security experts disseminated a set of serious security issues in social networks. In fact, the amount of personal information and sensitive data that may be managed by social networks make them an ideal target for hackers and hacking organizations. In this chapter, we will review the main social network cyberattacks and security issues reported since 2016. Our goal is to present a deep and accurate analysis of scenarios and results of these cyberattacks and security issues in order to determine a set of lessons learned and recommendations to enhance the security posture of the social networks.

2. MAIN REPORTED SOCIAL NETWORK CYBERATTACKS

In this section, we will classify the main social network cyberattacks and security issues into two categories : cyberattacks orchestrated against social networks, and security issues disseminated in social networks without any reported cyberattack.

2.1. Cyberattacks orchestrated against social networks :

The Dyn Cyberattack[1] and the Telegram[2] DDOS attacks :

On October 21st, 2016, a series of 3 Distributed Denial of Service (DDoS) attacks, also known as the 2016 Dyn Cyberattack [1], was reported. As the Dyn is the Domain Name System (DNS) provider of main social networks, this targeted and caused disruption to many of them.

A DDoS attack is a malicious attempt to disrupt normal traffic of a targeted service, network or server by flooding the target or its surrounding infrastructure with internet traffic. A DNS server is a system that is like a telephone book, that translates the web address, the name, into an Internet Protocol (IP) address, the telephone number. A DDoS attack on a DNS server can saturate the bandwidth and make the service unable to process legitimate traffic, therefore making itself unavailable [1].

These attacks reached a network traffic of 1 Tbps and affected social networks such as Twitter, Spotify, GitHub, Pinterest, Quora, Reddit, Slack,

among others. It was reported that the attacks were executed from tens of millions of Internet Protocol (IP) addresses at the same time, where each IP address is unique. This has been executed by a botnet consisting of many infected internet-connected or internet-of-things devices, such as IP cameras, DVRs and wireless routers, by a malware called Mirai [2].

In the same context, on June 12th, 2019, the popular encrypted messaging service Telegram was hit with a Distributed Denial of Service (DDoS) attack in Asia. Its chief executive, Parel Durov, said that Internet Protocol (IP) addresses were mostly coming from the People's Republic of China and experienced an all-state actor-sized DDoS of 200-400 Gb/s of junk, that coincided with events in the region.

Indeed, the increase of number of insecure connected devices has resulted in a surge of DDOS cyberattacks against internet services in general and social networks in particular. Thus, social networks infrastructures have to implement defenses measures to be prepared for prominent DDOS cyberattacks such as protection measures (e.g. packets and request filtering, machine learning based detection and network protection) and reaction measures (e.g. rerouting all traffic through a protectors' network) [3].

Instagram[3], Facebook[4],and Google+[5] data breaches owing to software vulnerabilities :

On August 30th, 2017, Instagram suffered a serious data breach with hackers gaining access to phone numbers and email addresses of 6 million verified users. Then, hackers had launched Doxagram, an online service, where users can search for stolen information related to one account for only $10. A data breach is a security incident where information is accessed without proper authorization [4]. This data breach was due to a bug in the mobile Application Programming Interface (API) that leaked information. The API was a set of functions and procedures that allows developers to access services from Instagram for their apps.

According to a security researcher from Kaspersky Labs[6], the issue resided in the password reset option in the mobile API, which exposed phone numbers and email addresses of the users.

In the same context, on September 28th, 2018, Facebook revealed that a few days prior, on September 25th, their engineering team discovered a security issue affecting almost 50 million accounts. Attackers exploited a vulnerability in Facebook's "View As", a feature that lets people see what

their profile looks like from someone else point of view. The vulnerability was a complex interaction of multiple issues in Facebook's code. In July 2017, they made changes to their video uploading feature.

This vulnerability allowed them to steal Facebook access tokens, from one account to another. Stolen Facebook access tokens could be used to take over people's accounts. Access tokens are digital keys that keep users logged into Facebook without the need to enter the password at each use of the app.

On December 10th, 2018, Google revealed a second bug, introduced in November during a previous platform update, in the Google+ Application Programming Interface (API) that could have been used to steal private data of nearly 52.5 million users. The bug resided in the Google+ People API endpoint that allowed apps, which were previously granted permission to view Google+ profile data, to incorrectly receive permission to view profile information that had been set to "not-public". More data were affected such as name, email address, occupation, age, skills, birthdate, nickname.

The data breaches of Google, Instagram and Facebook consolidate the need of effectively manage security risk during the software development and maintenance by being informed on new and evolving security threats, and continuously upgrading API and software systems to counteract and prevent them [5].

Quora[7] and Flipboard[8] data breaches owing to external unauthorized access :

On December 3rd, 2018, Quora revealed that 100 million of its users' data were compromised as a result of unauthorized access to one of their systems by a malicious third party. The compromised data contained name, email address, encrypted password, public and non-public content and actions and data imported from linked networks.

In the same context, on May 28th, 2019, Flipboard revealed that there was an unauthorized access to their databases, between June 2nd, 2018 and March 23rd, 2019, then April 21st and22nd, 2019, that leaked account information including names, usernames, email addresses and hashed passwords. In the Flipboard case, the unauthorized access was discovered after identifying suspicious activity in the environment where the databases reside.

Unauthorized access could be handled by the configuration of an accurate firewall to block unwanted inbound traffic to social network infrastructures. In addition, network-based and host-based intrusion detection and prevention system (NIDPS and HIDPS) can prevent intrusions and network mapping [9].

Taringa![9] data breaches owing to weak cryptographic protection:

On September 4th, 2017, LeakBase, a breach notification service, disclosed a hack where attackers allegedly stole records of nearly 29 million users of Taringa!, a popular Latin American social network. The stolen records contained usernames, email addresses and passwords protected with weak and outdated MD5 hashing algorithm [6].

A hashing process takes the original input and transforms it using a hashing algorithm, so that the output is unrecognizable [7]. MD5 is a popular but weak and outdated hashing algorithm, unless the bcrypt and scrypt algorithms, which are preferred to protect password over MD5 [8]. Since passwords were protected by MD5, the LeakBase team managed to crack 93.79% of the leaked passwords in days, cracking a total of nearly 27 million passwords.

Currently, older hashing algorithms can be brute forced in some seconds [10]. Thus, social network infrastructures must use only modern hashing algorithms like bcrypt and scrypt as they are secure to the brute forcing cyberattacks.

2.2. Security issues disseminated in social networks without any reported cyberattack:

Ello[11] [12], GitHub[13] [14], Twitter[15] [16], Facebook and Instagram[17] [18]: the security issues with the password requirements:

On November 11th, 2016, it was pointed out that social network Ello was displaying username and password strings visible in the Uniform Resource Locator (URL), or most known as web address, in plain text.

The issue resided in the form used to put the data in the URL. URLs should never contain sensitive information and if they do, there is a serious security problem as URLs would show up in the browsing history.

On May 1st, 2018, during a regular auditing, GitHub noted that a recently introduced bug in an in-house anti-spam system that logs metric from the password-reset form, exposed some user passwords in plain text in their secure internal logging system before finishing to perform bcrypt during a user-initiated password reset.

On May 3rd, 2018, after a regular auditing, Twitter asked all 300+ million users to reset their passwords due to a coding bug storing password in plain text. Concretely, the passwords were unmasked to a secure internal log only accessible by Twitter employees. In fact, Twitter must complete the hashing process of passwords, with state-of-the-art encryption technology such as bcrypt to mask a user's true password and make it difficult to read. A hashing process is a process that takes the original input and transforms it by using a hashing algorithm,so that the output is unrecognizable.

On March 21st, 2019, Krebs revealed that hundreds of millions of Facebook users had their account passwords stored in plain text and searchable by thousands of Facebook employees since 2012. In fact, a password should only be stored if it has gone through a hashing process that transforms it using an algorithm, so that the password is unrecognizable. On April 2019, it was estimated that millions of Instagram users were also affected.

Indeed, it is noticeable that all the major social networks had problems with secure password storage. Concretely, they must use a one-way cryptographic hash function to make it practically impossible to reverse the password[10]. In addition, it would be highly recommended for social networks to use a dynamic salt during the hashing process[11]. This means that for each password, a new salt is added before the hashing by a cryptographically strong random string generator[12][13].

Facebook[19], Google+[20], Twitter[21] [22] [23] and Instagram[24] [25] : the security issues related to software bugs :

From September 13th to September 25th, 2018, there was a bug in the Facebook photos Application Programming Interface (API). This API presented a set of functions and procedures that allows developers to access

services from Facebook for their apps. In this case, the API, which usually grants access to photos people share on their timeline, lets third-party developers view other photos, such as those shared on Marketplace or Facebook Stories, of up to 6.8 million Facebook users, shared or unpublished.

On October 8th, 2018, originally reported by The Wall Street Journal, ahead of Google's announcement on the matter, it was revealed that a software bug exposed the personal information of half a million users of its Google+social network. The exposed personal information included the name, email address, occupation, gender, age, as well as when this data was listed as private, as opposed to public. Google closed the bug in March 2018 shortly after learning of its existence, but it affected an Application Programming Interface (API) accessible to hundreds of developers. An API is a set of functions and procedures that allows developers to access services from Google+ for their apps. The bug was active between 2015 and 2018.

On December 1st, 2018, a French security researcher found a vulnerability in WordPress Plugin, "Social Network Tabs". The poorly written widget, last updated in 2013 and downloaded more than 53 thousand times, leaked Twitter "access_token", "access_token_secret", "consumer_key" and "consumer_secret" from their users by reading a PHP file "dcwp_twitter.php". Exposed keys had "read/write" access that can allow anyone to take control over Twitter accounts. Using the Common Vulnerabilities and Exposures (CVE), MITRE assigned the vulnerability CVE-2018-20555[26].

On September 12th, 2019, Zak Doffman revealed on Forbes that a newly discovered security vulnerability may have put Instagram, owned by Facebook, data at risk. It included users' real names, Instagram account number and full phone numbers.

This can be possible in two steps. First, the attacker uses a simple algorithm to brute force Instagram's login form, checking one phone number each time. The form replays if itis a valid number or not. A single instance of the algorithm can harvest more than 1,000 Instagram numbers per day. On average, 15,000 requests give 1,000 live numbers. Secondly, the attacker finds the account name and number linked to the phone number. This is possible using Instagram's Sync Contacts feature. A bot sets up a new account, and Instagram asks the new user whether they want to sync contacts. This typically returns a mass of account numbers and names, but without the ability to link these account details to telephone numbers.

As social networks are using web technologies, it is highly recommended to perform a daily web security scan in order to detect new software issues before discovering security vulnerabilities or security bugs for hackers [14]. A web security scan may reveal outdated server software, insecure cookie settings [15], SQL and Scripts Injection [16] [17], etc. In addition, the security vulnerability databases such as the National Vulnerability Database[27] need to be visited regularly to identify newly discovered security vulnerabilities within the technology used to implement the social networks.

WeChat[28] [29] [30], 63red Safe app[31], Instagram[32], and Facebook[33] [34] : the security issues related to weak security controls :

On March 4th, 2019, it was reported that 18 open and non-secure MongoDB databases of 364 million records from Chinese users were searchable on the internet. Each record contained ID numbers, photos, addresses, GPS location data, public and private conversations, file exchanges and information on the type of device being used. A messed-up firewall configuration left the database exposed and without security. A firewall is a network security device that monitors and controls all the incoming and outgoing network traffic. As social networks may involve millions or billions of connected users, it is remarkable to note that a more restrictive security firewall configuration would result in less performance [19]. Thus, the tradeoff between security and performance have to be considered [18].

On March 12th, 2019, according to a French security researcher who connects to the internet under the pseudonym of Elliot Anderson, the 63red Safe application, an app on local business reviews and talks, is leaking all its data through the server Application Programming Interface (API), which is left exposed online without authentication. The data includes usernames, emails, avatars, followers and following counts and profile creations or last updates All of this could be accessed by anyone who would look in the application's source code, use the API and extract data from the application's server with no difficulty or restriction. The API could also allow hackers to block users, tamper the database and hide unauthorized intrusions. If a social network uses an API as an external resource that can access or manage protected resource requests by external users, it must be able to authenticate the external users and to manage authorization

pertaining to the authentication process [20]. Authentication techniques include the HTTP Basic Authentication [21], the API Key Authentication [22], the OAuth Authentication [23], etc.

On May 20th, 2019, TechCrunch revealed that a massive database of 350,000 records of Instagram influencers, celebrities and brand accounts, owned by Mumbai-based social media marketing firm Chtrbox and hosted by Amazon Web Services (AWS), was exposed online without a password. When dealing with users' data, the database needed to be secured with a password. Each record contained users' public and private information such as bio, profile picture, number of followers, if they are verified, location by city and country, email address and phone number. Moreover, on September 4th, 2019, on TechCrunch, Zack Whittaker revealed that multiple databases across several geographies were found to be unprotected with passwords. With more than 419 million records, each of them contained the user's phone number, Facebook account ID, name, gender and location by country. When dealing with users' data, databases need to be protected with passwords. Unknown, is the owner of those databases as it is not Facebook's servers. Indeed, social networks have to ensure the best practices to safeguard the integrity and confidentiality of their databases [24]. The best practices include ensuring physical database security, implementing web application and database firewalls, managing database access tightly, auditing and monitoring database activity, etc.

3. CONCLUSION

After reviewing more than 20 data breaches and security issues discovered in all the main social network platforms and involving the personal data of hundreds of millions of users, it is interesting to note that the intuitive belief surrounding the implementation of data protection best practices relationship to technical maturity of these platforms, does not always hold in information security. The results reveal that a significant number of basic security issues were observed in all the frequently used social networks. We also observed that for the period of 2016-2019, almost all the frequently used social networks fail to put in place appropriate technical measures and controls to implement the data protection principles. Indeed, as the main business of social networks is to handle personal data, it must be designed and implemented by using the highest security measures and best data protection practices..

REFERENCES :

[1] Doshi, Rohan, Noah Apthorpe and Nick Feamster. "Machine Learning DDoS Detection for Consumer Internet of Things Devices", *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 29-35. IEEE, 2018.

[2] Kolias, Constantinos, Georgios Kambourakis, Angelos Stavrou and Jeffrey Voas. "DDoS in the IoT : Mirai and Other Botnets", *Computer* 50, nº 7 (2017) : 80-84.

[3] Wisthoff, Michael. "Ddos countermeasures", In *Information Technology-New Generations*, pp. 915-919. Springer, Cham, 2018.

[4] Manworren, Nathan, Joshua Letwat and Olivia Daily. "Why you Should Care About the Target Data Breach",*Business Horizons* 59, nº 3 (2016) : 257-266.

[5] Wisthoff, Michael. "DDoS Countermeasures," *Information Technology-New Generations*, pp. 915-919. Springer, Cham, 2018.

[6] Ah Kioon, Mary Cindy, Zhao Shun Wang and Shubra Deb Das. "Security Analysis of MD5 Algorithm in Password Storage", *Applied Mechanics and Materials*, vol. 347, pp. 2706-2711. Trans Tech Publications Ltd, 2013.

[7] Gupta, Piyush and Sandeep Kumar. "A Comparative Analysis of SHA and MD5 Algorithm", *Architecture 1* (2014) : 5.

[8] Sriramya, P. and R. A. Karthika. "Providing Password Security by Salted Password Hashing using Bcrypt Algorithm", *ARPN Journal of Engineering and Applied Sciences 10*, nº 13 (2015) : 5551-5556.

[9] Bul'ajoul, Waleed, Anne James and Mandeep Pannu. "Improving Network Intrusion Detection System Performance Through Quality of Service Configuration and Parallel Technology", *Journal of Computer and System Sciences 81*, nº 6 (2015) : 981-999.

[10] Klein, Daniel V. "Foiling the Cracker : A Survey of, and Improvements to, Password Security", *Proceedings of the 2nd USENIX Security Workshop*, pp. 5-14. 1990.

[11] Gauravaram, Praveen. "Security Analysis of Salt Password Hashes",*2012 International Conference on Advanced Computer Science Applications and Technologies* (ACSAT), pp. 25-30. IEEE, 2012.

[12] Sriramya, P. and R. A. Karthika. "Providing Password Security by Salted Password Hashing Using Bcrypt Algorithm",*ARPN Journal of Engineering and Applied Sciences 10*, nº 13 (2015) : 5551-5556.

[13] Boonkrong, Sirapat and Chaowalit Somboonpattanakit. "Dynamic Salt Generation and Placement for Secure Password Storing", *IAENG International Journal of Computer Science 43*, nº 1 (2016) : 27-36.

[14] Stuttard, Dafydd and Marcus Pinto. "The Web Application Hacker's Handbook : Finding and Exploiting Security Flaws", John Wiley & Sons, 2011.

[15] Khu-Smith, Vorapranee and Chris Mitchell. "*Enhancing the Security of cookies*", *International Conference on Information Security and Cryptology*, pp. 132-145, Springer, Berlin, Heidelberg, 2001.

[16] Halfond, William G., Jeremy Viegas and Alessandro Orso. "A classification of SQL-Injection Attacks and Countermeasures", *Proceedings of the IEEE international Symposium on Secure Software Engineering*, vol. 1, pp. 13-15. IEEE, 2006.

[17] Gupta, Shashank and B. B. Gupta. "Automated Discovery of JavaScript Code Injection Attacks in PHP Web Applications,"*Procedia Computer Science 78* (2016) : 82-87.

[18] Lyu, Michael R. and Lorrien KY Lau. "Firewall Security : Policies, Testing and Performance Evaluation", *Proceedings of 24th Annual International Computer Software and Applications Conference,* COMPSAC2000, pp. 116-121. IEEE, 2000.

[19] Salah, Khaled, Khalid Elbadawi and Raouf Boutaba. "Performance Modeling and Analysis of Network Firewalls", *IEEE Transactions on Network and Service Management 9*, nº 1 (2011) : 12-21.

[20] Corner, Mark D. and Brian D. Noble. "Protecting Applications with Transient Authentication", *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, pp. 57-70. 2003.

[21] Franks, John, Phillip Hallam-Baker, Jeffrey Hostetler, Scott Lawrence, Paul Leach, Ari Luotonen and Lawrence Stewart."HTTP Authentication : Basic and Digest Access Authentication" (1999) : 78.

[22] Heiland, Randy, Scott Koranda, Suresh Marru, Marlon Pierce and Von Welch. "Authentication and Authorization Considerations for a Multi-Tenant Service", *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure : Research, Experience, Applications and Models*, pp. 29-35. 2015.

[23] Leiba, Barry. "OAuth Web Authorization Protocol", *IEEE Internet Computing 16*, nº 1 (2012) : 74-77.

[24] Bertino, Elisa and Ravi Sandhu."Database Security-Concepts, Approaches, and Challenges", *IEEE Transactions on Dependable and Secure Computing 2*, nº 1 (2005) : 2-19.

[1] Report on Internet of Things Attacks since 2016

[2] https ://techcrunch.com/2019/06/12/telegram-faces-ddos-attack-in-china-again/

[3] https ://thehackernews.com/2017/08/instagram-breach.html

[4] https ://newsroom.fb.com/news/2018/09/security-update/

[5] https ://www.zdnet.com/article/google-hit-by-second-api-bug-impacting-52-5-million-users/

[6] https ://thehackernews.com/2017/09/instagram-hack-doxagram.html

[7] https ://www.quora.com/q/quora/Quora-Security-Update

[8] https ://about.flipboard.com/support-information-incident-may-2019/ ?noredirect=en_US

[9] https ://www.tripwire.com/state-of-security/latest-security-news/over-28-million-taringa-user-records-exposed-in-data-breach/

[10] https ://www.novatec-gmbh.de/en/blog/choosing-right-hashing-algorithm-slowness/

[11] https ://nakedsecurity.sophos.com/2016/11/21/alternative-social-network-ello-in-plain-text-password-glitch/

[12] https ://bestsecuritysearch.com/ello-plaintext-glitch-surprised-users/

[13] https ://www.zdnet.com/article/github-says-bug-exposed-account-passwords/

[14] https ://twitter.com/sanjuanswan/status/992457926999773184

[15] https ://krebsonsecurity.com/2018/05/twitter-to-all-users-change-your-password-now/

[16] https ://arstechnica.com/information-technology/2018/05/twitter-advises-users-to-reset-passwords-after-bug-posts-passwords-to-internal-log/

[17] https ://krebsonsecurity.com/2019/03/facebook-stored-hundreds-of-millions-of-user-passwords-in-plain-text-for-years/

[18] https ://newsroom.fb.com/news/2019/03/keeping-passwords-secure/

[19] https://developers.facebook.com/blog/post/2018/12/14/notifying-our-developer-ecosystem-about-a-photo-api-bug/

[20] https://www.theverge.com/2018/10/8/17951914/google-plus-data-breach-exposed-user-profile-information-privacy-not-disclosed

[21] https://techcrunch.com/2019/01/17/wordpress-plugin-leaked-twitter-account-access-tokens/

[22] https://twitter.com/fs0c131y/status/1085828186708066304

[23] https://github.com/fs0c131y/CVE-2018-20555#cve-2018-20555

[24] https://www.forbes.com/sites/zakdoffman/2019/09/12/new-instagram-hack-exclusive-facebook-confirms-user-accounts-and-phone-numbers-at-risk/#7fd4c07a2200

[25] https://au.finance.yahoo.com/news/instagram-security-flaw-phone-number-exposed-050722650.html

[26] https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-20555

[27] https://nvd.nist.gov/

[28] https://www.theverge.com/2019/3/4/18250474/chinese-messages-millions-wechat-qq-yy-data-breach-police

[29] https://www.bleepingcomputer.com/news/security/open-mongodb-databases-expose-chinese-surveillance-data/

[30] https://www.scmp.com/tech/enterprises/article/2188662/data-leak-exposes-364-million-chinese-social-media-profiles-tracked

[31] https://www.zdnet.com/article/yelp-for-conservatives-maga-app-leaks-users-data/

[32] https://techcrunch.com/2019/05/20/instagram-influencer-celebrity-accounts-scraped/

[33] https://www.forbes.com/sites/daveywinder/2019/09/05/facebook-security-snafu-exposes-419-million-user-phone-numbers/#643807a81ab7

[34] https://techcrunch.com/2019/09/04/facebook-phone-numbers-exposed/

2

LA SÉCURITÉ UTILISABLE : ENTRE L'INTERACTION HOMME-MACHINE ET LA SÉCURITÉ DE L'INFORMATION

Hervé Saint-Louis

Hervé Saint-Louis est professeur adjoint en médias émergents à l'Université du Québec à Chicoutimi. Chercheur en interaction homme-machine et en droit de l'information, il détient un doctorat de l'Université de Toronto. Il est aussi un animateur-bédéiste.

RÉSUMÉ

L'authentification, la vie privée, la sécurité des courriels, et les approches de sécurité contre les hameçonnages sont des sujets importants étudiés depuis des décennies par les chercheurs et les professionnels de la sécurité de l'information. Plusieurs des approches utilisées pour étudier ces phénomènes et d'autres touchant la sécurité de l'information passent par des analyses des systèmes informatiques, des réseaux, des infrastructures, des politiques, réglementations, standards, et décisions qui ont pour but la minimisation de risques. Or, la sécurité utilisable, une autre perspective qui considère plutôt le côté humain de toute interaction entre une personne et un système technique. Cette perspective basée sur les théories, les pratiques et les moyens d'évaluation du domaine de l'interaction homme-machine considère que la sécurité des systèmes informatiques et de l'information passe par une meilleure compréhension des utilisateurs des techniques. L'humain n'a jamais été oublié par les chercheurs et professionnels de la sécurité de l'information, cependant, il n'est pas toujours placé au cœur des préoccupations de ces derniers comme il l'est par les chercheurs en sécurité utilisable. Nous avons choisi d'utiliser le terme sécurité utilisable pour introduire le domaine intitulé « *usable security and privacy* » par nos collègues anglophones. Le chapitre qui suit est une introduction à ce champ d'étude multidisciplinaire en français se penchant sur l'histoire de ce domaine d'étude dynamique, il explique les échanges entre la sécurité utilisable, la sécurité de l'information, et ses composantes. De plus, le chapitre présente un cas étudiant l'inscription d'un nouveau compte Twitter en utilisant une analyse perceptuelle pour comprendre le point de vue d'un usager.

INTRODUCTION

Pour bien saisir le défi pratique et théorique de la sécurité utilisable, je commence cet article par trois problèmes classiques. Premier cas : les médecins d'un hôpital doivent changer le mot de passe des outils numériques de diagnostiques et de suivi de leurs patients mensuellement par mesure de sécurité. Cette condition a été imposée par le service des techniques[1] de l'information de l'hôpital. Pour contourner cette exigence, les médecins recyclent le mot de passe chaque mois en ne faisant que quelques modifications superficielles. Deuxième cas ; une plateforme de clavardage sur téléphone cellulaire requiert un mot de passe contenant des chiffres, des lettres minuscules et majuscules, des signes spéciaux et de plus, rejette toute configuration jugée trop « faible ». Ces conditions frustrent la plupart des usagers de l'application qui peinent à se souvenir du mot de passe qui doit être entré chaque fois qu'une session est démarrée. Troisième cas ; un système d'opération demande aux usagers de prendre des décisions majeures sur l'accès réservé à un logiciel de pare-feu installé sur l'ordinateur en utilisant des termes complexes qui ne sont pas familiers de la majorité des usagers. Certaines personnes acceptent tout ce qui est proposé sans trop comprendre et d'autres décident de cesser d'utiliser le pare-feu. Elles préfèrent naviguer sur des réseaux internes ou externes sans protection au lieu d'essayer de déchiffrer les indications du système d'exploitation et du pare-feu.

Les trois cas cités ci-dessus illustrent le type de cas traités par la sécurité utilisable. C'est un domaine de recherche situé entre l'informatique, les sciences sociales, les sciences cognitives et le design qui s'oriente sur l'usager pour explorer les défis reliés à la sécurité de l'information. Peut-être inconnu de certains, ce champ d'études a pris forme vers 1999.

Un constat primordial de ce domaine, quelquefois partagé par d'autres experts en sécurité (Schneier, 2015) est que la plus grande barrière à la sécurité de l'information est la personne en interaction avec une technique de l'information et des communications (TIC). On dit de cette personne qu'elle fait des erreurs (Norman, 2013) affectant l'efficacité des mesures de sécurité. La sécurité utilisable est empreinte d'une dialectique entre la

1. Dans ce chapitre, par besoin de justesse étymologique, je favorise le terme « technique » pour nommer ce que nous entendons habituellement par la technologie. En français, la technologie est l'étude des techniques. En utilisant le terme technique, j'évite un calque de la forme anglaise du mot « *technology*. »

sécurité et l'utilisabilité (Cranor & Garfinkel, 2004). Plus on augmente la sécurité des TIC, plus on risque d'affecter négativement l'utilisabilité pour l'usager (Renaud, 2004). Cependant, favoriser l'utilisabilité au détriment de la sécurité pour augmenter la facilité d'interaction avec des TIC peut créer des risques sérieux pour les usagers (Schultz, Proctor, Lien, & Salvendy, 2001).

On peut considérer l'interaction homme-machine (IHM) comme la discipline mère de la sécurité utilisable. Les problèmes d'interaction reliés à la vie privée, incluant la confidentialité, font partie de la sécurité utilisable. Les thèmes sur la vie privée dans la sécurité utilisable se démarquent au point que sécurité et vie privée sont considérées comme égales dans le domaine. La place de la vie privée dans la sécurité utilisable est d'ailleurs étudiée en détail ci-dessous. L'authentification, la sécurité des courriels, et les approches de sécurité contre les hameçonnages sont d'autres sous-domaines importants de la sécurité utilisable.

Plusieurs des approches utilisées pour étudier ces sous-domaines et autres problèmes touchant la sécurité de l'information passent par des analyses des systèmes informatiques, des réseaux, des infrastructures, des politiques, des réglementations, des standards, et des décisions juridiques qui ont pour but la minimisation de risques. Or, la sécurité utilisable considère plutôt le côté humain de toute interaction entre une personne et une technique. Une des orientations de la sécurité utilisable inspirée de l'IHM est de cesser de blâmer l'usager en favorisant le retrait des contraintes à l'utilisation de la sécurité de l'information par l'individu (Sasse, Brostoff, & Weirich, 2001).

Cette perspective basée sur les théories, les pratiques, et les moyens d'évaluation du domaine de l'IHM considère que la sécurité des systèmes informatiques et de l'information passe par une meilleure compréhension des usagers des TIC. L'humain n'a jamais été oublié par les chercheurs et professionnels de la sécurité de l'information (Edgar & Manz, 2017), cependant, son interaction avec les TIC n'est pas toujours placée au cœur des préoccupations des experts en sécurité.

D'autres termes tels que la sécurité conviviale ont été considérés pour introduire le domaine intitulé « *usable security and privacy* » par nos collègues anglophones, mais les traductions françaises de certains standards en la matière tels qu'ISO/TR 16982 :2002 utilisé dans le contexte de l'ergonomie de l'IHM font déjà référence au terme utilisabilité (ISO/TC 159/SC 4, 2005).

Les pages qui suivent sont une introduction à ce champ de recherche dynamique multidisciplinaire en français.

En premier lieu, j'introduis l'IHM et explique les valeurs de design que l'on peut utiliser pour analyser la sécurité et l'approche perceptuelle utilisée pour comprendre les phases historiques de la discipline. Deuxièmement, j'explore l'histoire de la sécurité de l'information en utilisant l'approche perceptuelle introduite précédemment. Troisièmement, en me basant sur l'approche perceptuelle, je donne un exemple d'évaluation du processus d'enregistrement à Twitter avec une optique de sécurité utilisable. Dans le cas présenté, j'évalue les défis et embûches à la vie privée auxquels un nouvel usager fait face en voulant créer un compte sur Twitter. L'emphase sur la confidentialité et la vie privée sert aussi d'exemple pour démontrer comment ces sujets influencent de plus en plus les enjeux face à la sécurité de l'information dans le domaine de la sécurité utilisable.

L'INTERACTION HOMME-MACHINE

L'interaction homme-machine est un domaine de recherche qui puise ses sources de plusieurs disciplines telles que les sciences de l'information, l'informatique, l'ingénierie, l'ergonomie, la psychologie, le design industriel (Grudin, 2012), l'anthropologie, et la sociologie. Dans le monde francophone, certains préfèrent utiliser le terme interaction humain-machine, d'autres, interaction personne-machine. Le terme original utilisant le générique « homme » pour l'être humain est encore la forme la plus répandue. Dans ce chapitre je continuerai l'utilisation de cette forme sans préjudice aux questions du genre humain.

La sécurité en IHM n'est qu'un des obstacles qui doit être affronté par les concepteurs de TIC. Dès 1985, Shneiderman (2016) concevait une liste de règles à respecter dans la conception d'interface en IHM. Ses règles (la cohérence, l'utilisabilité universelle, une réponse à toute interaction, l'indication de l'achèvement d'une action, la prévention des erreurs, les actions réversibles, le contrôle de l'usager, et une limite sur la charge cognitive demandée à la personne) (Shneiderman, 2016) peuvent être utilisées comme guides pour évaluer la sécurité, et d'autres aspects du design des TIC ciblant des valeurs humaines (Friedman & Kahn Jr., 2002).

En donnant l'exemple de quelques valeurs qui s'immiscent dans la conception des TIC (le bien-être humain, la vie privée, l'absence de parti

pris, la confiance, l'autonomie, etc.)[2] Friedman, Kahn, et Borning (2008), nomment les grands axes philosophiques que l'on peut opérationnaliser en IHM avec des règles telles que celles nommées par Shneiderman. C'est ainsi que l'on peut entrevoir la première contribution à la sécurité de l'information à partir de l'IHM.

En pratique, l'IHM se base sur les observations ethnographiques des usagers ou sur des expériences testant les interactions de participants avec des TIC. Des approches qualitatives basées sur des entrevues, des groupes focus, et des questionnaires sont souvent utilisées pour mieux comprendre l'usager en plus de l'observation des pratiques des personnes avec les TIC. De ce fait, l'IHM emprunte beaucoup de ses pratiques aux sciences cognitives. La compréhension des modèles mentaux des usagers qui permet de mieux envisager comment ils imaginent leurs interactions avec des TIC (Norman, 1983) est aussi très développée. Souvent, les chercheurs en IHM testent l'utilisabilité des interfaces de TIC avec des participants et raffinent les prototypes progressivement. Certaines mesures quantitatives comme l'analyse de Fitts (MacKenzie, 1992) ont été utilisées pour quantifier l'interaction entre l'humain et la machine. Une autre approche en utilisabilité est le test AB (Nielsen, 2005) où deux interfaces sont suggérées à des participants pour déterminer quels éléments de chaque version sont les mieux adaptés. Toutes ces approches sont reprises en sécurité utilisable.

Pour certains, il faut tracer l'origine de l'IHM avec les débuts de l'informatique même, et parfois, en amont. Dès 1945, le scientifique Vannevar Bush (1945) suggérait une machine capable de gérer l'information et de faire des calculs avancés pour servir les besoins de l'usager. L'emphase de plusieurs penseurs en information tels que Bush, Paul Otlet, ou H.G. Wells, était de faciliter l'usage de l'information par l'individu (Grudin, 2017).

Outre l'approche où l'on envisage la technique au service de l'humain, il faut entrevoir les TIC comme des extensions de l'individu (Licklider, 1960). Cette approche se base sur l'histoire de l'interaction en passant par les dispositifs d'interaction souris, écran, clavier, crayon électronique, etc.). On parle alors d'invention marquante tel que le Sketchpad inventé en 1963, la souris inventée en 1963, ou l'interface Xerox Star inventée en 1981. Au fur et à mesure, les besoins ergonomiques des usagers ont été pris en compte

2. Il peut y avoir des chevauchements des règles de Shneiderman avec les valeurs de Friedman, Kahn, et Borning. Par exemple, l'utilisabilité universelle est une idée présente dans les deux modèles.

et les modalités d'interaction (touché, vision, ouïe, etc.) se sont améliorées (Mackenzie, 2013).

Cette approche est utilisée pour comprendre l'ergonomie et les facteurs humains qui proviennent de l'ingénierie aéronautique. Les premières motivations des ingénieurs étaient le développement de cabines de pilotage améliorées pour réduire les risques d'accident par les pilotes d'avion durant la Deuxième Guerre mondiale (Vincente, 2003). Similairement, certains chercheurs tels que Nielsen (1995) étudiaient les fautes de design et d'ergonomie des TIC pour suggérer des améliorations et des standards qui devraient être suivis par les concepteurs et les développeurs de sites web et de logiciels. L'amélioration continue des modalités d'interaction est la base fondatrice des approches historiques pour comprendre l'histoire de l'IHM.

Dourish (2001) propose une troisième approche-cadre en cinq phases pour comprendre l'IHM. La première phase était électrique. L'ordinateur était une machine analogique constituée de composants électroniques à usage unique. Ses programmes n'étaient pas numériques, mais des artefacts physiques créés à l'extérieur et insérés dans la mémoire de l'ordinateur (Dourish, 2001, pp. 5-6). La deuxième phase était symbolique. Les humains interagissaient avec les ordinateurs via des codes alphanumériques qui résumaient le langage machine numérique des ordinateurs (Dourish, 2001, p. 7).

La troisième phase était textuelle. Les humains interagissaient avec des ordinateurs à l'aide de terminaux télétype et vidéo (Dourish, 2001, p. 9). La quatrième phase, qui a débuté durant les années 1980, était graphique. Des interfaces graphiques avec des icônes complétaient les interactions symboliques et textuelles permettant aux usagers de gérer les informations via l'écran (Dourish, 2001, p. 11).

Pour Dourish, l'informatique tangible et sociale est la prochaine phase de l'interaction humaine (2001). À l'époque, l'attention de Dourish était centrée sur le téléphone cellulaire et des techniques émergentes omniprésentes telles que des installations numériques interactives. Cette phase, toujours en évolution, tente de séparer l'interaction centrée sur l'écran en considérant d'autres TIC et différentes modalités d'interaction autre que le visuel pour inclure l'internet-des-objets, les drones, les voitures autonomes.

La cinquième phase de l'approche de Dourish tient compte des TIC autres que l'ordinateur. Elle rapproche l'IHM avec les facteurs humains et l'ergonomie qui sont des domaines où le design industriel a toujours été considéré. Pour certains, l'appellation anglaise « *human-computer interaction* » est trop restreinte pour la discipline, car en se concentrant sur l'ordinateur comme dispositif central d'interaction, elle omet la variété d'interaction avec toutes sortes de dispositifs avec lesquels l'humain interagit (Rogers, Preece, & Sharp, 2011).

Un thème qui rejoint tous les champs d'expertise de l'IHM qui peuvent sembler lointains l'un de l'autre est l'expérience usager. L'expérience usager est un champ émergent en IHM qui est souvent perçu épistémologiquement comme une forme améliorée de l'utilisabilité (Sauro et Lewis 2012 ; Norman 2013 ; Tullis et Albert 2013) ou comme un sous-ensemble de cette dernière (Weir, Douglas, Richardson, & Jack, 2010). Toutefois, Hassenzahl (2008), conçoit l'expérience usager comme un concept fondé sur la phénoménologie. La perception et le contexte de la personne sont pris en compte dans l'évaluation des interactions entre humains et techniques.

L'expérience usager tient compte des perceptions des usagers. Dourish (2001) explique l'histoire de l'IHM en tant que relations perceptuelles d'abord matérialisées, puis en passant à des formes plus abstraites de perception et d'interaction avant de revenir vers le physique comme on peut le constater dans la cinquième phase qui se base sur des objets intelligents.

HISTOIRE PERCEPTUELLE DE LA SÉCURITÉ DE L'INFORMATION

Je m'inspire de l'approche perceptuelle de Dourish pour expliquer la sécurité de l'information et son interaction avec l'humain.

Le droit américain définit la sécurité de l'information comme étant la protection de l'information des systèmes informatiques pour contrer les accès non autorisés, l'utilisation, le partage, la disruption, la modification et la destruction (44 USC § 3542 - Definitions, s.d.). Le cadre théorique utilisé par les autorités américaines et plusieurs experts en sécurité de l'information est le modèle intitulé en anglais *CIA*, soit le maintien de la

confidentialité, l'intégrité, la disponibilité[3] de l'information par des mesures de protection (44 USC § 3542 - Definitions, s.d.). Ce modèle a été critiqué par Parker (1998, p. 6) comme étant limité et en particulier, incapable de protéger l'information et les systèmes informatiques des facteurs humains. Parker propose des dimensions supplémentaires pour mieux évaluer et protéger l'information (1998), mais l'usage de ses propositions n'est pas adopté ou aussi répandu dans les pratiques des experts en sécurité (Andress, 2011, p. 6).

Parker ajoute les dimensions authenticité, l'utilité, et du contrôle au modèle CIA de la sécurité (Parker, 1998, p. 240). Il s'intéresse à l'utilité, la véracité de l'information retenue et la possession de cette dernière. Bien que Parker ajoute des dimensions nécessaires à la sécurité de l'information, ces ajouts ne sont pas orientés vers l'usager, quoi qu'en dise l'expert dans ses arguments en faveur de la compréhension des facteurs humains (Parker, 1998, p. 6).

Yost (2007) affirme que l'établissement des normes de sécurité en information et en informatique n'a jamais été l'objectif de l'armée américaine qui s'est vue dans l'obligation de développer des standards rapidement pour permettre l'interopérabilité entre les équipements de combat et les ressources des grands ordinateurs. Ce qui était recherché était la facilité d'utilisation structurelle et organisationnelle.

La facilité d'interopérabilité révèle une dimension humaine de l'information où la commodité d'utilisation primait sur la sécurité. Les quelques usagers informatiques des grands ordinateurs des années 1940 et 1950 bénéficiaient de ce que l'on appelle la sécurité par l'obscurité. Quand un nombre restreint de personnes peuvent interagir avec l'information et les systèmes informatiques, la rareté d'usagers confère une forme de sécurité indéniable.

D'ailleurs, Yost écrit que la sécurité de l'information n'existait pratiquement pas parce que peu d'opérateurs avaient accès aux grands ordinateurs du milieu du XXe siècle (2007, p. 600). Tout comme les Chinois du 18e siècle qui, selon l'historien du renseignement David Kahn, n'ont pas réussi à mettre au point des mesures cryptographiques adéquates parce

3. En anglais, le terme disponibilité est intitulé « *availability* ». J'ai préféré utiliser « *disponibilité* » au lieu du terme accessibilité pour la traduction en français.

que très peu de Chinois savaient lire (Kahn, 1996, p. 74), la sécurité de l'information dans les premiers ordinateurs se faisait par l'obscurité.

Cependant, les mesures d'obscurité ne pouvaient plus satisfaire les besoins de sécurité dans les années 1960 et 1970 en raison des débuts de l'informatique partagée. L'informatique partagée a permis aux équipes de plusieurs d'utiliser les ressources d'un ordinateur, telles que la même base de données ou la même bibliothèque, pour une application utilisée par multiples usagers du même ordinateur (Saltzer & Schroeder, 1975). Yost affirme par exemple que l'informatique partagée qui augmentait le niveau d'interaction usager et les risques de sécurité dans l'armée avait conduit à la création de systèmes de classification tels que top secret, secret, confidentiel et non classifié (2007, p. 604).

Alors que l'ère de l'informatique personnelle persistait, des risques de sécurité externes persistaient, mais les erreurs survenues lors des interactions entre usagers constituaient des risques importants pour la sécurité de l'information. La connaissance de l'utilisabilité et de l'informatique personnelle a joué un rôle dans le type d'erreurs commises par les usagers. Même si plusieurs usagers pouvaient toujours utiliser le même ordinateur personnel, le risque concernait moins les données centralisées et les niveaux d'accès des différentes parties.

LA SÉCURITÉ UTILISABLE ?

Si la sécurité de l'information a pris sa place dans l'esprit des chercheurs, il faut se rappeler que les questions portant sur l'utilisabilité des mesures de protection des systèmes informatiques brillaient par leur obscurité. D'après Garfinkel et Ritcher Lipford (2014), peu de recherches entre 1970 et 1990 considéraient cette thématique.

Garfinkel et Ritcher Lipford (2014) attribuent la première mention d'un besoin de rendre la sécurité de l'information acceptable et utilisable pour l'usager à Saltzer et Schroeder (1975). Saltzer et Schroeder, dans leur article sur les mesures de protection de l'information en informatique, mentionnent que l'acceptation psychologique est nécessaire dans la conception des interfaces informatiques pour rendre facile l'usage de mécanismes de protection correctement pour l'usager (1975).

Saltzer et Schroeder (1975) mentionnent même le concept des modèles mentaux qui figure amplement dans la littérature de l'IHM (Ackerman &

Greutmann, 1990 ; Caroll & Anderson, 1987 ; Genther & Grudin, 1996 ; Norman, 2013 ; Sasse, 1997 ; Staggers & Norcio, 1993) pour justifier la corrélation des designs des mécanismes de protection avec les attentes des usagers.

La première recherche importante en sécurité utilisable est celle de Karat (1989) qui a effectué des tests d'ergonomie usager dans le cadre de l'implantation d'un nouveau logiciel de sécurité sur un processeur central utilisé par des usagers de systèmes informatiques IBM. Les travaux de Karat consistent à appliquer les méthodes et les pratiques de l'IHM à un outil de sécurité de l'information, traitant ce dernier de la même façon que l'on traite tout autre outil informatique.

Ce que les recherches de Karat ont clairement démontré c'est que les usagers de systèmes informatiques de sécurité pouvaient avoir les mêmes frustrations que les autres usagers (1989). La deuxième situation que Karat démontre est que la sécurité peut créer une barrière à l'ergonomie et l'utilisation d'un système informatique (1989). L'application sur le processeur central original forçait les usagers à s'authentifier continuellement dès qu'il avait un changement de fonction et d'opération. En se basant sur des mesures classiques des estimations des interfaces usagers, telles que le taux de complétion d'une tâche, le temps nécessaire pour compléter une tâche, et la performance sans erreur, Karat a évalué la sécurité de l'information dans le contexte où elle est utilisée par les individus.

Le thème des barrières dans l'usage usuel d'usager est le sujet de « *Why Johnny Can't Encrypt* » (2005) de Whitten et Tygar. Cet article est devenu un modèle pour une série de recherches montrant les limites d'utilisation des systèmes de sécurité destinés à protéger les usagers. Il propose aussi une charte normative pour évaluer la sécurité des TIC. Une série d'articles (Herzberg, 2009 ; Leon, Ur, Balebako, et Cranor, 2012 ; Olejnik, Castelluccia, & Janc, 2012 ; Sheng, Broderick, Koranda, & Hyland, 2006) reprend le thème « *Why Johnny can't…* », l'appliquant a toute sorte de problèmes en sécurité utilisable depuis lors. Cet article fait aussi partie des classiques de la littérature de la sécurité utilisable axée sur la démonstration des limites des pratiques de sécurité de l'information pour les usagers. Comme mentionné ci-dessus, « *Users Are Not the Enemy* » (1999) d'Adams et Sasse ont étudié la manière dont les usagers créaient des systèmes pour contourner les mesures de sécurité des systèmes d'authentification d'entreprise.

La focalisation sur la démonstration des limites des systèmes de sécurité existants avec les usagers a conduit une grande partie de la recherche en sécurité utilisable à se concentrer sur l'authentification. L'authentification est importante, car elle gère la protection des usagers en interaction avec les TIC manipulant les données personnelles des gens. Ces recherches peuvent porter sur l'évaluation comparative des principaux systèmes d'authentification (Bonneau, Herley, van Oorschot, & Stajano, 2012) ou sur des systèmes spécifiques tels que la biométrie (Coventry, 2005), les mots de passe graphiques (Monrose & Reiter, 2005) ou même les captchas (Yan & El Ahmad, 2008).

Les autres domaines d'intérêt pour les chercheurs en sécurité utilisable comprennent les courriels (Sheng, Broderick, Koranda, & Hyland, 2006), la messagerie (Gaw, 2009), et le cryptage de cette dernière (Renaud, Volkamer, & Renkema-Padmos, 2014). Le domaine de recherche le plus distinct de la discipline est celui de la protection de la vie privée. La recherche sur la sécurité utilisable liée à la protection de la vie privée est devenue suffisamment importante pour constituer un domaine de préoccupation égal à la sécurité dans des forums de recherche spécialisés tels que le Symposium sur la sécurité et la vie privée utilisables (SOUPS). La recherche sur la protection de la vie privée éloigne la sécurité utilisable de ses origines purement instrumentales et commence à répondre aux préoccupations liées à la manière dont les personnes interagissent avec la technique dans l'économie de l'information.

Nissenbaum (2004) explique qu'il y a trois dimensions à la vie privée qui se déploient selon le contexte. Le premier est basé sur les limites à l'intrusion de l'état dans la vie des gens (Nissenbaum, 2004). Le pouvoir de l'état est réglementé pour empêcher ce dernier d'avoir un droit de regard absolu sur la vie des gens. Le deuxième vise la protection de l'information personnelle qui peut être collectée sans égard pour qui effectue la collecte (Nissenbaum, 2004). Cette dimension vise la protection des secrets des gens. Finalement, Nissenbaum (2004) nous présente une autre dimension de la vie privée basée sur l'espace personnel d'un individu, tel, sa maison, son ordinateur personnel, son téléphone cellulaire.

Le plan théorique que nous fournit Nissenbaum nous offre une vision globale des phénomènes reliés à la vie privée sans entrer directement dans ce que je qualifie de la relation sujet-objet qui caractérise l'utilisation de TIC par des usagers. La dimension de la vie privée qui touche aux protections envers l'état a moins d'importance en sécurité utilisable, car le lieu du

problème n'est plus structurel et social. Il est individuel et basé sur des actions portées par un individu avec les TIC. La protection de l'information personnelle collectée par support médiatique ainsi que l'espace non physique revendiqué par les TIC nous ramène vers le personnel. La protection de la vie privée se complique quand l'implication de l'IHM est considérée.

Certains auteurs regardent plus attentivement l'espace personnel et l'interaction entre l'humain et la technique. Culnan (2000), définit la protection de la vie privée comme le contrôle que les individus exercent sur leurs informations personnelles. Ackerman et Mainwaring (2005, pp. 382-383), utilisant la définition de Culnan comme point de départ, décrivent la vie privée comme étant subjectivement individuelle et socialement positionnée. Les individus perçoivent la vie privée différemment en fonction de l'application et du contexte d'utilisation. Par exemple, ils font valoir que les usagers perçoivent la vie privée différemment lorsqu'ils utilisent des services bancaires personnels et des médias sociaux (Ackerman & Mainwaring, 2005, p. 383). Les usagers peuvent percevoir leurs informations comme confidentielles lorsqu'ils utilisent un système bancaire personnel. Sur un site de média social, les usagers peuvent se sentir plus libres de partager leurs informations publiquement. Brunk décrit la définition de la vie privée donnée par le chercheur Eli Noam comme « le lieu où les droits d'information de différentes parties se rencontrent (2005, p. 402). »

La protection de la vie privée n'est cependant pas la seule préoccupation de protection des architectes de l'information au stade de la conception. Plusieurs spécialistes considèrent la protection de la vie privée comme un élément de la sécurité (Bonneau, et autres, 2012 ; Mihajlov, Josimovski et Jerman-Blazič, 2011). Tout en développant un cadre d'évaluation de la sécurité utilisable dans les mécanismes d'authentification, Mihajlov, Josimovski et Jerman-Blazič (2011, p. 333) ont inclus la protection de la vie privée parmi les nombreux critères. Dans une étude similaire évaluant les avantages des méthodes d'authentification alternatives en termes d'utilisabilité, de déploiement et de sécurité, Bonneau et autres (2012, p. 5) décrivent la protection de la vie privée comme un élément de la sécurité.

Cela semble contredire le cadre habituel où l'on perçoit la sécurité comme une restriction sur la vie privée, en particulier depuis les évènements du 11 septembre 2001, où les sociétés sont en proie à la surveillance et au contrôle de l'information (Bambauer 2013 ; Deibert 2012). La nature de la sécurité sur laquelle j'enquête est pertinente pour les individus, par

opposition aux états et aux organisations. Il s'agit de la sécurité personnelle des personnes qui inclut leur vie privée lorsqu'elles interagissent avec des TIC. Toutefois, dans la pratique, lorsque les TIC conservent des informations personnelles sur des personnes, ils le font de manière confidentielle et avec l'accord tacite ou explicite des usagers (Siegel, 1979). Par conséquent, il est plus approprié de dire que c'est la confidentialité qui est protégée plutôt que la vie privée.

ANALYSE DE MÉDIAS SOCIAUX PERCEPTUELLE

L'analyse perceptuelle du processus d'enregistrement d'un compte sur Twitter présenté ci-dessous démontre, d'une part, pourquoi étudier et comprendre la sécurité utilisable, et offre, d'autre part, plusieurs pistes pour traiter de problèmes dépassant souvent l'informatique et sa sécurité. L'analyse est qualitative et peut facilement être refaite avec plusieurs participants dans une étude de cas expérientielle si l'objectif est de tester comment les usagers surmontent les embûches lors de l'enregistrement d'un compte sur Twitter. L'IHM a la particularité d'emprunter beaucoup des approches expérientielles de la psychologie et autres sciences cognitives. L'évaluation qualitative présentée ci-dessus servait à comprendre et répertorier le processus d'enregistrement d'un compte avant de tester le tout dans une expérience avec des participants.

Les politiques de confidentialité, souvent les premiers documents sur lesquels les usagers interagissent lors de l'utilisation de sites et d'applications, les informent de la manière dont les données les concernant sont collectées, puis, utilisées. (Cranor L. F., 2005, p. 448). Cependant, la compréhension des usagers et les interactions avec ces documents sont souvent problématiques (Jensen & Potts, 2004). Pour les usagers, ils peuvent sembler complexes, souvent ignorés ou convenus sans une lecture attentive (Milne & Culnan, 2004).

Les recherches consacrées à la compréhension de l'interaction des usagers avec les politiques de confidentialité se sont concentrées sur leur contenu (Grossklags & Good, 2007), leurs modèles mentaux (Coopamootoo & Groß, 2014) ou des questionnaires et des entrevues sur les perceptions de participants (Adams & Sasse, 1999). Bien que ces approches centrées sur l'usager aient produit des résultats, je propose une autre perspective axée sur l'évaluation des modèles de conception des opérateurs de plateforme qui définissent l'interaction de l'usager avec les politiques de

confidentialité et de sécurité. Au lieu de simplement passer en revue les textes de ces politiques ou d'interroger des participants sur leurs impressions, je propose une approche inspirée de l'approche perceptuelle de Dourish (2001).

Mon évaluation perceptuelle nous donne un aperçu de la manière dont les politiques de confidentialité peuvent affecter les usagers, mais au lieu d'enquêter sur leur contenu, nous souhaitons comprendre comment elles sont conçues pour les interactions entre usagers. Les analyses de la politique de confidentialité des usagers examinent souvent comment les usagers perçoivent et lisent les documents (Jensen & Potts, 2004), leurs stratégies pour traiter les problèmes de confidentialité et la manière dont ils définissent leurs paramètres personnels pour réduire les risques (Johnson, Egelman, & Bellovin, 2012).

Cette approche produit des résultats, mais ne se concentre pas sur le contexte dans lequel les usagers interagissent avec les plateformes. Les usagers ont également la possibilité d'interagir avec les politiques de confidentialité, en particulier lorsqu'ils créent de nouveaux comptes sur des plateformes. Souvent, lorsqu'ils créent un compte pour une plateforme, les usagers doivent accepter le contenu d'une politique de confidentialité ou les conditions d'utilisation correspondantes.

C'est là que les évaluations perceptuelles des politiques de confidentialité deviennent les plus pertinentes. Ces évaluations permettent aux chercheurs de cartographier et de comprendre ce qui se passe après que l'usager a sauté au bas de l'accord ou ignoré une invitation à ouvrir un lien distinct pour prendre connaissance du contenu de la politique et de la façon dont leurs informations sont collectées par les plateformes.

Dans cette section, j'ai effectué une évaluation approfondie des politiques de confidentialité publiée par Twitter du 25 mai 2018 (Twitter, Inc, 2018) telles qu'elles sont présentées aux usagers finaux interagissant avec les plateformes. En particulier, nous observons comment les usagers sont informés des règles de confidentialité lorsqu'ils s'enregistrent pour de nouveaux comptes sur les plateformes. L'enregistrement est l'un des moments les plus importants où les usagers interagissent avec les politiques de confidentialité. L'autre site d'interaction avec les politiques de confidentialité est l'endroit où les usagers ajustent leurs paramètres de confidentialité. Les paramètres de confidentialité sont développés à partir des politiques de confidentialité.

ANALYSE PERCEPTUELLE DE TWITTER

L'authentification et l'enregistrement sur Twitter permettent à l'usager de devenir à la fois diffuseur et public. La vérification de l'identité de l'usager doit permettre l'accès à cette plateforme interactive. Cependant, la situation de marchandisation en cours sur Twitter oblige la société à rassembler des informations sur ses usagers.

Lors de l'enregistrement d'un nouveau compte, le nouvel usager est confronté à un menu lui demandant de saisir d'abord son nom complet, puis un numéro de téléphone ou une adresse courriel. Twitter demandant aux nouveaux usagers d'enregistrer leur numéro de téléphone ou leur adresse électronique trahit la double origine de sa plateforme de communication, où messages textes et Internet sont des lieux équivalents pour les usagers qui souhaitent diffuser leurs communications.

Sur Twitter, les noms des usagers sont traités différemment du sobriquet utilisé dans les messages. Les usagers ont une double identité. Le sobriquet est destiné à la messagerie tandis que le nom est utilisé pour identifier l'usager. Au lieu d'être un élément de données caché, comme sur Facebook ou basé sur une adresse électronique, telle que Google, le surnom (également appelé descripteur) est le moyen utilisé par les usagers pour accéder à la plateforme. Chaque interaction avec les autres dans le flux se fait par le surnom.

La page d'inscription affiche des liens vers les conditions d'utilisation, la politique de confidentialité et les règles de Twitter en matière d'utilisation des témoins. Ce sont des documents complets auxquels l'usager peut facilement accéder avant de saisir des données personnelles sur la page d'inscription. C'est un meilleur affichage des politiques que Facebook, qui oblige les usagers à accepter un contrat et à ajuster leurs paramètres après s'être enregistrés, ou à Google qui supprime tous les autres éléments de l'interface et la navigation de l'usager, ce qui l'oblige à faire défiler une page pour accepter les conditions antérieures avant d'aller de l'avant.

Cependant, il reste des problèmes avec la page d'enregistrement de Twitter. Il existe tout d'abord une option présélectionnée permettant à Twitter de proposer aux personnes inscrites des annonces personnalisées pour leurs comptes. Une autre option présélectionnée propose à l'inscrit de laisser Twitter suivre toute son interaction avec des contenus de la plateforme à travers le web. Par défaut, ces options sont cochées. Encore une fois, si les inscrits veulent comprendre en quoi consistent les annonces

personnalisées et le suivi sur le web, ils doivent accéder à une page différente et lire un document de politique de vie privée.

Une valeur d'utilisation secondaire pour les usagers plus enracinés est la possibilité d'avoir un public prêt pour leurs tweets. Gagner des abonnés est un processus ludique sur Twitter où les usagers avec le plus large auditoire gagnent en statut et en influence. Sur Facebook et Google, le nombre d'amis et de contacts d'un usager a son importance, mais ce n'est pas un symbole de statut comme c'est le cas sur Twitter.

Le processus de création d'un nouveau surnom sur Twitter est plus difficile que sur Facebook ou Google. Sur Facebook, en suivant la stratégie de nom réel, plusieurs usagers peuvent partager le même nom. L'identifiant qui les sépare sur la plateforme de Facebook n'est pas leur nom. Sur Google, il existe également des difficultés lors de la tentative de recherche de nouveaux noms. Tout comme les noms de domaine, tous les bons noms sont déjà pris.

L'architecture de Twitter ne prend en charge que les caractères de soulignement et principalement les lettres de l'alphabet latin. Les usagers doivent faire preuve de stratégie quant à la longueur de leur surnom pour avoir des noms viables pouvant facilement être inclus dans les interactions avec le réseau. Il y a d'ailleurs une limite de caractères pour un surnom. Twitter utilise un autre identifiant associé au compte d'un usager, car les sobriquets peuvent être modifiés.

Le processus de génération d'un pseudonyme Twitter est une étape importante dans la socialisation du nouvel usager Twitter avec la plateforme. Les identités sont valorisées sur Twitter. Les usagers sont encouragés à remplir des biographies abrégées d'eux-mêmes pour se tailler une petite propriété et pour projeter leur personnalité sur la plateforme. L'identité devient un moyen pour l'usager Twitter d'annoncer son canal de diffusion à d'autres usagers. En quelques mots, elle doit faire appel à un large public et l'inciter à la suivre, augmentant ainsi la portée de ses propres messages.

RÉSUMÉ DE L'ÉVALUATION PERCEPTUELLE

Twitter offre aux usagers la possibilité de consulter la politique telle quelle, sans aucune modification. Ayant clairement exposé sa politique de confidentialité, Twitter vise à aider les nouveaux usagers à naviguer et à se familiariser avec sa plateforme. Il présente ses fonctionnalités et tente de

minimiser la difficulté d'obtenir un pseudonyme Twitter propre et unique. La présentation de la politique de confidentialité devient une nécessité à ne pas manquer, offerte par l'opérateur de la plateforme, mais ne fait pas partie intégrante de son expérience usager.

Twitter tente de rassembler autant d'informations personnelles que possible des nouveaux usagers avant qu'ils n'utilisent le site. Ce faisant, Twitter oblige les usagers à interagir avec d'autres plateformes et systèmes, ce qui rend leur interaction sociale et dépendante d'autres sites d'interaction. Cette pratique est conforme à ce que Dourish décrit comme une informatique sociale (2001). L'informatique sociale consiste à interagir socialement avec plusieurs techniques dans le cadre d'une activité. Par exemple, les usagers de Twitter sont invités à interagir avec leurs courriels et leurs applications de contact. Cela oblige les usagers à synthétiser des informations sur eux-mêmes et sur d'autres personnes grâce à diverses techniques interconnectées. Lorsque les usagers ajustent leurs paramètres de confidentialité dans un environnement ludique, l'activité qu'ils effectuent est identique à celle du jeu, mais dans un contexte différent.

RECOMMANDATIONS

Twitter devrait offrir des options paramétrables permettant aux usagers de contrôler leurs options de confidentialité. La confidentialité et les paramètres de sécurité doivent être proposés par Twitter. Avec des paramètres de confidentialité, ils contrôleraient davantage que les données échangées entre les usagers et certains tiers. De tels contrôles permettraient aux usagers de décider globalement de la quantité de données qu'ils choisiraient de partager en permanence avec la plateforme. Par exemple, les usagers doivent pouvoir désactiver les mesures comportementales utilisées par les plateformes pour suivre leur interaction.

L'objectif final de l'enregistrement des usagers est probablement de ne pas perdre de temps à lire les conditions de confidentialité et de sécurité des services lorsqu'ils rejoignent une plateforme. L'authentification et l'enregistrement deviennent des moyens d'accéder à un domaine technologique. Twitter devrait demander un accès aux données en cas de besoin. Le principe de ne demander que les données nécessaires au besoin est une pratique éthique de la protection que l'on retrouve dans des pratiques tel que le « *Privacy-by-Design* » de Cavokian (2009).

L'étude de cas ci-dessus est un exemple assez typique d'une étude en sécurité utilisable qui diffère beaucoup des méthodes d'évaluations utilisées en sécurité de l'information.

CONCLUSION

La sécurité utilisable est un domaine de recherche connexe à la sécurité de l'information qui puise ses approches méthodologiques de l'IHM. L'usager, qu'il soit un néophyte en sécurité (Whitten & Tygar, 2005) ou un expert (Chiasson, Biddle, & Somayaji, 2007) mérite un environnement d'interaction utilisable et basé sur des normes de design sécuritaire. La vie privée est devenue au cours du temps une des composantes les plus importantes de ce domaine de recherche et partage même l'intitulé de SOUPS, le forum de recherche le plus important sur ce sujet. L'emphase sur la vie privée reflète une plus grande prépondérance des questions liées à la surveillance par les états ou les grandes entreprises dans le domaine de l'information, comme on peut le voir dans l'étude perceptuelle sur l'authentification et l'enregistrement de nouveau compte sur Twitter. La sécurité utilisable recentre le problème de la sécurité sur l'individu en essayant de prévenir toute embuche tout en facilitant l'utilisation et en améliorant l'expérience de ce dernier avec les techniques.

BIBLIOGRAPHIE

44 USC § 3542 - Definitions. (s.d.). Consulté le 26 novembre 2012, sur *Legal information Institute - Cornell University Law School* : http ://www.law.cornell.edu/uscode/text/44/3542

Ackerman, M. S., & Mainwaring, S. D. (2005). "Privacy Issues and Human-Computer Interaction". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Security and Usability : Designing Secure Systems that People Can Use* (pp. 381-399), Sebastapol : O'Reilly.

Ackermann, D., & Greutmann, T. (1990). "Experimental Reconstruction and Simulation of Mental Models". Dans D. Ackermann, & M. J. Tauber (Éds.), *Mental Models and Human-Computer Interaction 1* (pp. 136-150). Amsterdam : North-Holland.

Adams, A., & Sasse, M. A. (1999). "Users Are Not the Enemy". *Communications of the ACM, 42*(12), 40-46.

Andress, J. (2011). "The Basics of Information Security : Understanding the Fundamentals of InfoSec in Theory and Practice". Waltham, Massachusetts, United States of America : Syngress.

Bambauer, D. E. (2013). "Privacy versus Security". *Journal of Criminal Law & Criminology, 103*(3), 667-683.

Bonneau, J., Herley, C., van Oorschot, P. C., & Stajano, F. (2012). The Quest to Replace Passwords : A Framework for Comparative Evaluation of Web Authentication Schemes. *2012 IEEE Symposium on Security and Privacy* (pp. 553 - 567). San Francisco : IEEE.

Brunk, B. (2005). "A User-Centric Privacy Space Framework". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Security and Usability : Designing Secure Systems that People Can Use* (pp. 401-420). Sebastapol : O'Reilly.

Bush, V. (1945, July). "As We May Think". *The Atlantic*. Récupéré sur https ://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Carroll, J. M., & Anderson, N. S. (1987). "Mental Models in Human-Computer Interaction : Research Issues About What the User of Software Knows'. Washington, D.C. : National Academy Press.

Cavoukian, A. (2009). "Privacy by Design". Toronto : Information and Privacy Commissioner of Ontario.

Chiasson, S., Biddle, R., & Somayaji, A. (2007). "Even Experts Deserve Usable Security : Design Guidelines for Security Management Systems". *Workshop on Usable IT Security Management - Symposium on Usable Privacy and Security*, (pp. 1-4). Pittsburgh.

Coopamootoo, K. P., & Groß, T. (2014). "Mental Models for Usable Privacy : A Position Paper". *Human Aspects of Information Security, Privacy, and Trust : Second International Conference.* Heraklion.

Coventry, L. (2005). "Usable Biometrics". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Usable Security : Designing Secure Systems that People Can Use* (pp. 175-197). Sebastopol : O'Reilly.

Cranor, L. F. (2005). "Privacy and Privacy Preferences". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Security and Usability : Designing Secure Systems that People Can Use* (pp. 447-471). Sebastapol : O'Reilly.

Cranor, L. F., & Garfinkel, S. (2004). "Secure or Usable ?". *The IEEE Computer Society*, 16-18.

Culnan, M. J. (2000). "Protecting Privacy Online : Is Self-Regulation Working ?". *Journal of Public Policy & Marketing, 19*(1), 20-26.

Deibert, R. (2012, November). The Growing Dark Side of Cyberspace (. . . and What To Do About It)". *Penn State Journal of Law & International Affairs, 1*(2), 260-274.

Dourish, P. (2001). "Where the Action Is : The Foundations of Embodied Interaction". Cambridge : MIT Press.

Edgar, T. W., & Manz, D. O. (2017). "Research Methods for Cyber Security'. Cambridge : Syngress.

Friedman, B., & Kahn Jr., P. H. (2002). Human Values, Ethics, and Design. Dans A. Sears, & J. A. Jacko (Éds.), *The Human-Computer Interaction Hhandbook* (pp. 1177-1201). Abingdon.

Friedman, B., Kahn Jr., P. H., & Borning, A. (2008). "Value Sensitive Design and Information Systems". Dans K. E. Himma, & H. T. Tavani (Éds.), *The Handbook of Information and Computer Ethics* (pp. 69-101). Hoboken : Wiley.

Garfinkel, S., & Ritcher Lipford, H. (2014). "Usable Security : History, Themes, and Challenges". Williston, Vermount : Morgan & Claypool Publishers.

Gaw, S. (2009, June). "Ideals and Reality : Adopting Secure Technologies and Developing Secure Habits to Prevent Message Disclosure". Princeton, New Jersey : Princeton University.

Grossklags, J., & Good, N. (2007). "Empirical Studies on Software Notices to Inform Policy Makers and Usability Designers". *Financial Cryptography and Data Security.* 341355.

Grudin, J. (2012). "A Moving Target : The Evolution of Human-Computer Interaction". Dans J. A. Jacko (Éd.). *The Human-Computer Interaction Handbook : Fundamentals, Evolving Technologies, and Emerging Applications* (éd. 3, pp. xxvii-lxi). CRC Press : Boca Raton.

Grudin, J. (2017). "From Tool to Partner : The Evolution of Human-Computer Interaction". Williston, Vermont : Morgan and Claypool.

Hassenzahl, M. (2008). "User experience (UX) : Towards an Experiential Perspective on Product Quality". *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine* (pp. 11-15). Metz : ACM.

Herzberg, A. (2009). "Why Johnny cCcan't Surf (safely) ? Attacks and Defenses for Web Users". *Computers & Security*, pp. 63-71.

ISO/TC 159/SC 4. (2005). « ISO/TR 16982 :2002 Ergonomie de l'interaction homme-système - Méthodes d'utilisabilité pour la conception centrée sur l'opérateur humain ». Genève : International Organization for Standardization.

Jensen, C., & Potts, C. (2004). "Privacy Policies as Decision-Making Tools : An Evaluation of Online Privacy Notices". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 471-478). Vienna.

Johnson, M., Egelman, S., & Bellovin, S. M. (2012). "Facebook and Privacy : It's Complicated". *SOUPS '12 Proceedings of the Eighth Symposium on Usable Privacy and Security*. Washington, D.C.

Kahn, D. (1996). "The Codebreakers". New York : Scribner.

Karat, C.-M. (1989). "Iterative Usability Testing of a Security Application". *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 273-277). Denver : Human Factors and Ergonomics Society.

Leon, P. G., Ur, B., Balebako, R., & Cranor, L. F. (2012). "Why Johnny Can't Opt Out : A Usability Evaluation of Tools to Limit Online Behavioral Advertising". *CHI*, (pp. 1-38).

Licklider, J. C. (1960). "Libraries of the future". Cambridge : MIT Press.

Mackenzie, I. S. (2013). "Human-Computer Interaction : An Empirical Research Perspective". Waltham, Massachusetts : Elsevier.

MacKenzie, S. (1992). "Fitts' Law as a Research and Design Tool in Human-Computer Interaction". *Human-Computer Interaction,* 7, pp. 91-139.

Mihajlov, M., Josimovski, S., & Jerman-Blazič, B. (2011). "A Conceptual Framework for Evaluating Usable Security in Authentication Mechanisms – Usability Perspectives". *2011 5th International Conference on Network and System Security*, (pp. 332-336).

Milne, G. R., & Culnan, M. J. (2004). "Strategies for Reducing Online Privacy Risks : Why Consumers Read (or Don't Read) Online Privacy Notices". *Journal of Interactive Marketing, 18*(3), 15-29.

Monrose, F., & Reiter, M. K. (2005). "Graphical Passwords". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Security and Usability : Designing Secure Systems that People Can Use,* (pp. 157-174). Sebastopol : O'Reilly.

Nielsen, J. (1995, 1er janvier). "*10 Usability Heuristics for User Interface Design*". Consulté le January 27 2015, sur http ://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. (2005, 15 août). "Putting A/B Testing in Its Place". Consulté le 11 juillet 2019, sur *Nielsen Norman Group* : https ://www.nngroup.com/articles/putting-ab-testing-in-its-place/

Nissenbaum, H. (2004). "Privacy as Contextual Integrity". *Washing Law Review Association*, *79*, pp. 101-139.

Norman, D. A. (1983). "Some Observations on Mental Models". Dans D. Gentner, & A. L. Stevens (Éds.), *Mental Models* (pp. 7-14). Hillsdale, New Jersey : Lawrence Erlbaum Associates, Inc.

Norman, D. A. (2013). "The Design of Everyday Things : Revised and Expanded Edition". New York : Basic Books.

Olejnik, L., Castelluccia, C., & Janc, A. (2012). "Why Johnny Can't Browse in Peace : On the Uniqueness of Web Browsing History Patterns". *5th Workshop on Hot Topics in Privacy Enhancing*, (pp. 1-16). Vigo.

Parker, D. B. (1998). "Fighting Computer Crime : A New Framework for Protecting Information". New York : Wiley Computer Publishing.

Renaud, K. (2004). "Quantifying the Quality of Web Authentication Mechanisms. A Usability Perspective". *Journal of Web Engineering, 3*(2), pp. 95-123.

Renaud, K., Volkamer, M., & Renkema-Padmos, A. (2014). "Why Doesn't Jane Protect Her Privacy ?" Dans E. De Cristofaro, & S. Murdoch (Éd.), *Privacy Enhancing Technologies. PETS 2014. Lecture Notes in Computer Science. 8555*, pp. 244-262. Cham : Springer.

Rogers, Y., Preece, J., & Sharp, H. (2011). "Interaction Design : Beyond Human-Computer Interaction". New York : Wiley.

Saltzer, J. H., & Schroeder, M. D. (1975). "The Protection of Information in Computer Systems". *Proceedings of the IEEE. 63*, pp. 1278-1308. Cambridge : ACM.

Sasse, M. A. (1997). "Eliciting and Describing Users' Models of Computer Systems". University of Birmingham, School of Computer Science. Birmingham : University of Birmingham.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). "Transforming the 'Weakest Link' — a Human/Computer Iinteraction Approach to Usable and Effective Security. *BT Technology Journal, 19*(32), 122–131.

Sauro, J., & Lewis, J. R. (2012). "Quantifying the User Experience : Practical Statistics for User Research (éd. 1st Edition)". Amsterdam : Morgan Kaufmann.

Schneier, B. (2015). "Secrets and Lies : Digital Security in a Networked World". Indianapolis : Wiley Publishing.

Schultz, E. E., Proctor, R. W., Lien, M.-C., & Salvendy, G. (2001). "Usability and Security An Appraisal of Usability Issues in Information Security Methods". *Information Security Methods Computers & Security, 20* (7), pp. 620–634.

Sheng, S., Broderick, L., Koranda, C., & Hyland, J. J. (2006, July 12–14). "Why Johnny Still Can't Encrypt : Evaluating the Usability of Email Encryption Software". *Symposium on Usable Privacy and Security*. Pittsburgh.

Shneiderman, B. (2016). "The Eight Golden Rules of Interface Design". Consulté le 5 juillet 2019, sur http ://www.cs.umd.edu/~ben/goldenrules.html

Siegel, M. (1979, April). "Privacy, Ethics, and Confidentiality". *Professional Psychology, 10* (2), pp. 249–258.

Staggers, N., & Norcio, A. (1993). "Mental Models : Concepts for Human-Computer Interaction Research". *International Journal of Man-Machine Studies* (38), pp. 587–605.

Tullis, T., & Albert, W. (2013). "Measuring the User Experience : Collecting, Analyzing, and Presenting Usability Metrics (éd. 2nd Edition)". Amsterdam : Morgan Kaufmann.

Twitter, Inc. (2018, mai 25). « Politique de confidentialité de Twitter ». Récupéré sur Twitter : https ://twitter.com/privacy ?lang=fr

Vincente, K. (2003). "The Human Factor". Toronto : Alfred A. Knopf Canada.

Weir, C. S., Douglas, G., Richardson, T., & Jack, M. (2010). "Usable Security : User Preferences for Authentication Methods in eBanking and the Effects of Experience". *Interacting with Computers* (22), pp. 153–164.

Whitten, A., & Tygar, J. (2005). "Why Johnny Can't Encrypt : A Usability Evaluation of PGP 5.0". Dans L. F. Cranor, & S. Garfinkel (Éds.), *Security and Usability* (pp. 669–692). Sebastopol : O'Reilly.

Yan, J., & El Ahmad, A. S. (2008). "Usability of CAPTCHAs or Usability Issues in CAPTCHA Design". *Proceedings of the 4th Symposium on Usable Privacy and Security* (pp. 44–52). New York : ACM.

Yost, J. R. (2007). "A History of Computer Security Standards". Dans K. d. Leeuw, & J. Bergstra (Éds.), *The History of Information Security : A Comprehensive Handbook* (pp. 595–621). Amsterdam : Elsevier.

3

PROTÉGER AUTANT NOS DONNÉES QUE LES FONDEMENTS DE LA DÉMOCRATIE EN PRIVILÉGIANT LES LOGICIELS LIBRES

Mathieu Gauthier-Pilote

Formé en conception de systèmes informatiques (1998-2001), Mathieu Gauthier-Pilote débute sa carrière comme responsable des technologies de l'information chez un éditeur québécois de jeux ludo-éducatifs. À compter de 2011, il est travailleur autonome dans l'industrie du numérique. Spécialisé dans les logiciels libres et les technologies du Web, il agit comme conseiller, administrateur de systèmes, développeur et formateur auprès de PME et d'OSBL. En 2012, il devient chargé des projets numériques à la Fondation Lionel-Groulx, un organisme de bienfaisance qui se consacre à la promotion de l'histoire du Québec. Dans le cadre de ce mandat, il offre notamment de la formation à Wikipédia. Parallèlement à sa carrière professionnelle, il est membre (2003-), administrateur (2012-) et président du conseil (2015-) de FACiL, un OSBL dont la mission est de promouvoir l'appropriation collective de l'informatique libre par les Québécoises et les Québécois. Pour cet organisme, il a rédigé plusieurs articles, présenté des mémoires et donné des conférences. En 2014, il a également monté un cours pour l'Upop Montréal intitulé « L'informatique libre : droits, libertés et bien commun dans le cyberespace ». Sa plus récente publication substantielle est un long article intitulé « La tête dans les nuages. Le secteur public face aux offres commerciales d'infonuagique publique », paru en mai 2019 dans la revue L'Action nationale.

RÉSUMÉ

Cet article explore quelques-uns des enjeux que la centralisation d'Internet par les GAFAM pose à la démocratie, ainsi que les réponses que le milieu du logiciel libre tente de leur apporter, particulièrement depuis l'électrochoc des révélations d'Edward Snowden, à l'été 2013. Sont abordés non seulement les enjeux de sécurité et de vie privée, mais plus

largement ceux de nos libertés et de nos droits fondamentaux, tout comme ceux de la distribution du pouvoir, de la connaissance et de la richesse dans la société numérique du 21e siècle. À travers des exemples concrets de protocoles, de plateformes et d'applications proposés par différentes communautés libristes, sont révélés les avantages des quatre libertés du logiciel libre, mais également les défis auxquels sont confrontés des solutions technologiques qui n'ont de force que si elles sont massivement adoptées par les internautes et pas uniquement en raison de leur « gratuité ».

INTRODUCTION

Il ne fait pas de doute que les spécialistes de la sécurité chez Google, Apple, Facebook, Amazon et Microsoft (GAFAM) travaillent très fort pour empêcher que des tiers malveillants accèdent illégalement aux données que nous leur avons confiées. Cependant, qui nous protège contre les usages de nos données qui sont peut-être légaux, mais pas forcément légitimes et respectueux de notre vie privée, et auxquels se livrent très certainement ces grandes entreprises installées au cœur de l'économie marchande de la donnée ? Plus généralement, que faisons-nous collectivement pour réduire la collecte massive des données permettant le pistage, le profilage et la surveillance permanente des internautes par les États et les géants du numérique, phénomène global gravissime qui menace nos droits à la vie privée, à la liberté d'expression, de communication, d'association ?

Depuis plus de 30 ans, la philosophie du logiciel libre nous enseigne que le domaine de liberté protégé par les licences des logiciels libres n'est pas seulement utile et bénéfique aux utilisateurs et aux utilisatrices d'appareils numériques, il est devenu *essentiel* à la préservation des libertés qui sont au fondement de la société démocratique. Il n'y a que dans le mouvement du logiciel libre qu'existe véritablement le souci de protéger les utilisateurs et les utilisatrices d'appareils numériques à la fois contre les filous qui cherchent à exploiter les failles de sécurité des systèmes numériques et aussi contre les abus et les erreurs des personnes et des entreprises qui conçoivent les logiciels qui traitent nos données.

En effet, ce n'est pas un bien grand secret que les GAFAM qui dominent l'industrie mondialisée du numérique sont presque toujours en conflit d'intérêt lorsqu'il est question des données des internautes : les modèles économiques choisis par ces entreprises reposent précisément sur la valorisation marchande de données comportementales produites grâce à

la collecte massive de données, qui ne devraient jamais être exploitées (ou même produites) si le mot « éthique » a encore un sens.

Rappelons aussi que depuis les révélations d'Edward Snowden à l'été 2013, il est clair pour tous les gens informés que le problème numéro un d'Internet est celui de la centralisation (infrastructures, équipements, données, expertise) et de ses conséquences néfastes sur nos droits fondamentaux et sur la redistribution du pouvoir, de la connaissance et de la richesse. Si les logiciels libres, les services décentralisés et distribués et le chiffrement de bout en bout font nécessairement partie de la solution à ce problème, leur adoption massive par les internautes n'est pas écrite dans le ciel et comme en toute chose, il faudra accepter de se battre sur les terrains politique, social, économique et culturel pour gagner.

Dans cet article, je vais développer quatre points : 1. les avantages du logiciel libre pour la sécurité informatique et la protection des utilisateurs et des utilisatrices d'appareils numériques contre les fonctionnalités malveillantes des logiciels peu éthiques ; 2. le problème de la centralisation d'Internet et ses conséquences néfastes ; 3. les nouveaux protocoles et les nouvelles plateformes et applications sociales – décentralisées ou distribuées – promues par le milieu du logiciel libre pour « réparer » Internet, reprendre le contrôle sur nos données, mieux protéger toutes les libertés nécessaires à la sauvegarde de la démocratie dans le 21e siècle numérique ; 4. quelques-unes des batailles qui nous attendent sur les terrains politique, social, économique et culturel pour sortir les applications sociales libres, éthiques, décentralisées et solidaires des marges où elles sont aujourd'hui en 2020.

1. LES AVANTAGES DU LOGICIEL LIBRE POUR LA SÉCURITÉ INFORMATIQUE ET LA PROTECTION DES UTILISATEURS ET UTILISATRICES D'APPAREILS NUMÉRIQUES CONTRE LES FONCTIONNALITÉS MALVEILLANTES

La cybersécurité est certainement une affaire très sérieuse pour les GAFAM : ces multinationales ont les moyens d'en faire une priorité et leur clientèle – les entreprises comme les particuliers – s'attend à rien de moins que l'excellence de leur part.

Ainsi, bien que des erreurs et des lacunes soient inévitables, comme le montrent les fuites de données massives qui font régulièrement sensation dans les grands médias, tout doit nous porter à croire qu'elles déploient de très grands efforts pour empêcher que des *tiers malveillants* accèdent

illégalement aux données que nous leur avons confiées. La préservation de leur réputation comme de leurs intérêts commerciaux les y incite fortement.

Malheureusement, les modèles d'affaires de ces multinationales, surtout à l'ère de l'infonuagique et des mégadonnées, les empêchent de faire le choix éthique qui serait – notamment en matière de sécurité et de vie privée – le plus à l'avantage de leurs clients et des internautes en général : celui du logiciel libre, celui de la protection des quatre libertés fondamentales des utilisateurs et des utilisatrices d'appareils numériques[1]. Ce choix, qui les conduirait à *exposer toutes les fonctionnalités* des logiciels qui collectent et traitent les données de leurs clients et des internautes, elles ne veulent pas le faire. Elles ne le veulent pas, car il donnerait aux utilisateurs et aux utilisatrices des moyens de protection contre toutes sortes d'abus de pouvoir et de violation de droits de la part des entreprises qui sont les propriétaires et les véritables maîtres des logiciels dont nous sommes très nombreux à dépendre quotidiennement.

Mais de quels types d'abus de pouvoir parle-t-on exactement ? Et comment est-il possible de se protéger contre eux ? Voyons cela avec un peu de détails.

1.1. Les fonctionnalités malveillantes et ce qu'elles révèlent sur le caractère politique du logiciel

Commençons par affirmer un fait mal connu : l'introduction de *fonctionnalités malveillantes* dans les logiciels par les principaux joueurs de l'industrie du numérique est monnaie courante. En 2019, on compte plus de 400 cas documentés de fonctionnalités malveillantes dans le répertoire en ligne de la Free Software Foundation (FSF)[2], qui en propose d'ailleurs une typologie très intéressante et pédagogique. Ainsi, les logiciels conçus par les entreprises qui n'adhèrent pas à l'éthique du logiciel libre – l'essentiel de l'industrie – contiennent bien souvent des fonctions de censure, de dissimulation, de tromperie, des portes dérobées, des menottes numériques,

1. « Qu'est-ce que le logiciel libre ? », *gnu.org*, juin 2019. https://www.gnu.org/philosophy/free-sw.html
2. La FSF est un organisme de bienfaisance américain fondé en 1985 par Richard Stallman, le père du mouvement pour le logiciel libre. Sa mission est de promouvoir les libertés et de défendre les droits des utilisateurs et des utilisatrices d'appareils numériques. https://www.fsf.org/about/

des incompatibilités volontaires, des cas délibérés d'insécurité, d'ingérence, de sabotage, de surveillance, de manipulation psychologique, etc.[3]

En explorant les nombreux cas du répertoire de la FSF, on prend rapidement conscience des importantes conséquences politiques et sociales qui se trament parfois derrière les décisions en apparence purement techniques prises par les développeurs et les développeuses lors de l'écriture du code source des logiciels. Pour bon nombre de ces décisions prises lors de la conception et du développement des logiciels, personne n'était là pour représenter les utilisateurs et les utilisatrices et défendre leurs libertés, leurs droits et leurs intérêts. Le parti qui avait le gros bout du bâton – le propriétaire qui paie le développement et récolte les profits – ne s'est pas gêné pour se donner tous les droits et se mettre en position d'abuser de son pouvoir.

1.2. Les libertés que les propriétaires de logiciels prennent à nos dépens

Les développeurs et les développeuses sont, règle générale, des employé·e·s, qui ont renoncé à leur droit d'auteur et signé un accord de non divulgation en échange du salaire ou du contrat que leur donne une entreprise. C'est cette entreprise qui détient tous les droits de propriété sur le fruit de leur travail. L'entreprise concède ensuite des droits d'utilisation restreints et révocables aux utilisateurs et aux utilisatrices de sa propriété au moyen d'une *licence de logiciel*, qui est essentiellement une liste d'interdits : copier, prêter, louer, concéder à un tiers, décompiler ou désassembler par rétroingénierie, contourner les restrictions techniques qui empêchent ou limitent certains usages, etc. C'est le modèle du logiciel non libre ou *privateur de liberté*[4], qui se développe avec la microinformatique dans les années 1970.

Dans ce système de relations entre le petit nombre des propriétaires des logiciels et le grand nombre des propriétaires d'appareils numériques que nous sommes tous, nous nous retrouvons dans l'*incapacité de savoir* ce que font vraiment les applications exécutées sur nos téléphones, nos

3. « Le logiciel privateur est souvent malveillant », *gnu.org*, juin 2019. https://www.gnu.org/proprietary/proprietary.fr.html

4. Expression française proposée par Richard Stallman pour bien rendre le sens de *proprietary software* (littéralement « *logiciel qui a un propriétaire* »).

tablettes, nos postes de travail, nos serveurs, etc. En effet, dérobé de la vue des utilisateurs et des utilisatrices, le code source d'un logiciel privateur de liberté s'écrit dans le secret et est par conséquent comparable à une boîte noire dont on ne sait rien du fonctionnement interne.

C'est le contraire avec le logiciel libre : ceux et celles qui écrivent le code source utilisent leur droit d'auteur pour protéger, par une licence, la liberté de tous et de toutes d'utiliser, de copier, d'adapter et de redistribuer, tel quel ou modifié, le fruit de leur travail.

Si l'on admet que des décisions importantes et souvent de nature politique se prennent lors de la conception et de l'écriture des logiciels qui commandent les appareils numériques qui sont aujourd'hui omniprésents dans nos vies, on comprend sans trop de difficulté pourquoi le domaine de liberté protégé par les licences de logiciel libre est devenu absolument nécessaire à la société de l'information dans laquelle nous vivons au 21e siècle.

Dans son essai de 1999 intitulé *Code and Other Laws of Cyberspace*, le juriste américain Lawrence Lessig est allé jusqu'à soutenir la thèse célèbre que le code – du logiciel et même du matériel informatique – est au cyberespace ce que la loi est à l'espace politique, soit l'incarnation d'une norme par laquelle il est possible de réguler la conduite des humains en société[5]. Si c'est vrai, alors il n'y a qu'un pas à faire pour conclure que le processus de fabrication des logiciels doit devenir « transparent », au sens qu'il doit être soumis aux mêmes principes démocratiques qui permettent aux citoyens et aux citoyennes des États de la planète de défendre leurs libertés, leurs droits et le bien commun. À moins bien sûr de renoncer entièrement à réguler le cyberespace, ce que Lessig refuse comme bien d'autres, l'auteur de ces lignes compris.

1.3. Une sécurité informatique plus complète et l'espoir de renverser la surveillance de masse

Tout le monde se déclare bien sûr pour la démocratie et pour la vertu, mais quels avantages *concrets* la liberté procure-t-elle en matière de sécurité informatique et plus généralement de cybersécurité ?

5. Voir *Code and Other Laws of Cyberspace* (Basick Books, 1999, 297 p.) ou la seconde édition *Code : Version 2.0* (Basic Books, 2006, 410 p.) publiée sous licence libre Creative Commons BY-SA 2.5 à l'adresse http ://codev2.cc

Notons d'abord que plusieurs des enjeux de sécurité informatique n'ont rien à voir avec le code source mais plutôt avec l'administration, la configuration et l'utilisation au quotidien des systèmes d'information par les humains. Ils sont par conséquent sensiblement les mêmes que le logiciel soit libre ou pas. Cependant, lorsqu'il est question de la *sécurité des applications*, les utilisateurs et les utilisatrices d'un logiciel libre sont en meilleure position que les autres, car ils peuvent soumettre toutes les fonctionnalités du logiciel à un *audit indépendant* dans leur propre intérêt, avec ou sans l'accord de son auteur. Ils peuvent également collaborer, voire mutualiser leurs ressources en la matière[6]. Ainsi, l'exercice par les utilisateurs et les utilisatrices de la liberté d'étudier le fonctionnement d'un logiciel libre permet-il de mieux se protéger contre les failles de sécurité au niveau du code source, qui sont souvent causées de manière accidentelle par l'ignorance, la négligence ou l'erreur humaine. L'exercice de cette liberté leur donne également la possibilité de se protéger contre les *fonctionnalités malveillantes* qu'un propriétaire de logiciel pourrait vouloir introduire dans son produit ou son service contre l'intérêt des utilisateurs et des utilisatrices, parfois au détriment de la sécurité, du droit à la vie privée ou des autres droits fondamentaux[7].

Ainsi, en n'adoptant pas l'éthique du logiciel libre, les GAFAM – comme la plupart des entreprises, grandes et petites de l'industrie du numérique – montrent qu'en matière de cybersécurité, elles sont plus intéressées par la protection de leur commerce et de leur réputation que par la protection de leurs clients contre les abus dont elles sont elles-mêmes la source ou le vecteur et qui doivent continuer d'être cachés de la vue du public en utilisant contre lui tous les ressorts du droit d'auteur. Elles ont en effet besoin de logiciels privateurs de liberté, développés en secret, pour continuer le pistage, le profilage et la surveillance permanente des internautes à des fins commerciales.

---

6. Ces libertés mises en commun peuvent être utilisées pour aller très loin dans l'autodéfense face à des adversaires potentiels, comme c'est le cas avec le projet des *reproductible builds*, auquel participent de nombreuses communautés de logiciels libres comme Debian, Arch, Fedora, FreeBSD, etc. Voir l'adresse https ://reproducible-builds.org

7. J'attire l'attention sur les cas *délibérés* d'abus de pouvoir des propriétaires des logiciels, mais je n'ignore pas non plus ces autres cas où les propriétaires sont contraints de mal agir par la *corruption* d'un agent externe.

2. LE PROBLÈME DE LA CENTRALISATION D'INTERNET ET DE SES CONSÉQUENCES NÉFASTES

Les avantages des logiciels libres qui découlent entièrement de la transparence inhérente à leur processus de fabrication, du libre accès à leur code source, sont considérables, mais ils ne règlent évidemment pas tous les problèmes de cybersécurité, notamment ceux qui découlent de la surveillance de masse. Il ne servirait par exemple pas à grand-chose de forcer la libération du code source des logiciels des GAFAM si nous devions en même temps conserver l'architecture *centralisée* des écosystèmes numériques que ces grandes entreprises ont bâtis avec les logiciels en question. Nous serions toujours captifs des services en ligne qu'elles ont déployés sur Internet pour collecter nos données, pister nos mouvements, analyser nos comportements.

Pour cesser d'être captifs, il faut bien sûr nous placer en position de savoir concrètement ce que font nos appareils numériques, mais il nous faut aussi user de notre liberté d'adapter les logiciels à nos besoins, voire d'en écrire de nouveaux lorsque c'est nécessaire. Ce dont nous avons un urgent besoin collectif si nous souhaitons mettre fin à la surveillance de masse des GAFAM c'est de nous attaquer concrètement à la centralisation d'Internet[8].

En effet, depuis les révélations d'Edward Snowden à l'été 2013, il est clair pour tous les gens renseignés que le problème numéro un d'Internet est celui de la centralisation (infrastructures, équipements, données, logiciels, expertise) et de ses conséquences néfastes pour nos droits fondamentaux comme pour la redistribution du pouvoir, de la connaissance et de la richesse

Qu'est-ce que cette *centralisation* et pourquoi est-elle à la source de quelques-unes des plus importantes pathologies de la société numérique, à commencer par la surveillance de masse ?

8. Naturellement, il faut aussi mettre fin à la surveillance de masse des agences de renseignement des États, facilitée et renforcée par celle des GAFAM. Des réponses politiques comme techno-sociales sont nécessaires dans les deux cas.

2.1. La topologie des réseaux informatiques : plus que des détails techniques

Lorsque les milieux de l'industrie et de la recherche en sécurité, des hackeurs et des libristes ou des militants des droits fondamentaux parlent de centralisation, ils font d'abord référence à une notion de topologie des réseaux informatiques. En effet, les réseaux peuvent être structurés de différentes manières, qu'il est convenu de ranger dans trois ensembles : centralisés, décentralisés et distribués.

Dans les réseaux centralisés, les participants au réseau (les clients) dépendent d'un nœud central (le serveur), tandis que dans ceux qui sont dits décentralisés, les nœuds centraux sont multiples et les participants peuvent conséquemment *choisir* le serveur qui aura leur préférence (par exemple le plus proche géographiquement, le plus performant, le mieux sécurisé, le moins cher, celui qui est communautaire, celui qui nous appartient, etc.). Dans les réseaux distribués, il n'y a pas de nœuds centraux du fait que tous les participants se distribuent les fonctions de serveur (tous les clients sont également des serveurs).

La *décentralisation* informatique est donc le processus par lequel on réduit la centralisation des systèmes de type client-serveur en multipliant les nœuds serveurs. La *distribution* du calcul et du stockage sur les réseaux, grâce à des systèmes pairs-à-pairs, va plus loin en éliminant la nécessité de passer par un tiers de confiance (un serveur).

En étudiant attentivement le fonctionnement et les cas d'utilisation réels des différentes architectures des réseaux, on découvre des enjeux de pouvoir, de contrôle, de hiérarchie, de confiance, qui ont forcément une dimension politique dans les sociétés humaines qui les conçoivent et les exploitent. Les hackeurs et les libristes, attachés aux fondements décentralisés d'Internet et du Web, sont horrifiés par les succès des géants du numérique à rendre populaires des applications et des services en ligne reposant sur des architectures centralisées qu'ils contrôlent largement. C'est presque sans exagération qu'au courant de la décennie 2010 on a

commencé à parler dans divers milieux technos de la mort d'Internet[9] (ou du Web[10]) et de la nécessité de le « réparer » ou de le « reconstruire »[11].

Dans le contexte du mouvement qui prône la redécentralisation d'Internet, l'objectif principal qui est visé est d'abord l'accroissement de la participation des internautes dans les prises de décision qui affectent leur vie numérique.

L'idée est que la multiplication des serveurs sur le réseau, implique la concurrence de fournisseurs de services, qui peuvent être des entreprises locales, des associations sans but lucratif, des coopératives, des fédérations régionales, nationales, internationales, bref, des entités qui répondent à des critères éthiques et des objectifs politiques que les entreprises multinationales rencontrent rarement.

2.2. La centralisation et ses effets néfastes sur la cybersécurité globale et sur la vie privée

La multiplication des fournisseurs de services en mode client-serveur et la participation d'un grand nombre d'internautes aux réseaux distribués offrent des avantages certains à plusieurs niveaux, notamment en matière de cybersécurité et de vie privée.

Sur le plan strict de la cybersécurité, l'argument en faveur de la décentralisation est très puissant : les grands hébergeurs de données et d'applications qui centralisent Internet constituent des *cibles de choix* pour les attaquants ! Lorsqu'on peut espérer accéder aux informations de 200 000 clients d'un coup, le jeu en vaut la chandelle. Par contre, si les informations des mêmes 200 000 clients sont éparpillées dans différents systèmes (à la maison, sur nos appareils, ou chez un fournisseur local de confiance, à but

9. Oram, Andy, "How did we end up with a centralized Internet for the NSA to mine ?", *O'Reilly Radar*, 8 janvier 2014. http://radar.oreilly.com/2014/01/how-did-we-end-up-with-a-centralized-internet-for-the-nsa-to-mine.html.

   Granick, Jennifer Stisa, « The End of the Internet Dream ? », *Wired*, 17 août 2015. https://www.wired.com/2015/08/the-end-of-the-internet-dream/

10. Staltz, André, "The Web Began Dying in 2014", *Staltz.com*, 30 octobre 2017. https://staltz.com/the-web-began-dying-in-2014-heres-how.html

11. *Reset The Net*, 5 juin 2014. https://www.resetthenet.org

    Kahle, Brewster, "Locking the Web Open : Rethinking the World Wide Web", *The Next Web*, 10 avril 2015. https://thenextweb.com/insider/2015/04/10/locking-the-web-open-why-we-need-to-rethink-the-world-wide-web/

lucratif ou sans but lucratif, de petite ou de moyenne taille) ou alors distribuées dans le réseau et protégées par du chiffrement de bout en bout, les cibles deviennent *diffuses* et l'attaque massive devient bien moins praticable ou carrément impraticable.

Ce n'est pas tout. Une migration réussie des données des internautes depuis des systèmes centralisés appartenant à des entreprises dont les modèles économiques reposent sur la collecte massive, le pistage et le profilage, vers des systèmes décentralisés appartenant à un ensemble hétéroclite d'organisations aux modèles économiques différents, doit forcément être avantageux en matière de protection de la vie privée et de la sécurité.

Il ne faut jamais perdre de vue que les données personnelles les plus « sécuritaires » sont naturellement celles qui ne sont jamais ni produites, ni collectées, ni stockées. Viennent ensuite celles qu'on doit produire, mais qu'on n'est pas obligé de stocker ou de faire circuler sur le réseau. Finalement, en bout de ligne, il y a les autres données personnelles forcément stockées quelque part et qu'on peut être tenu de communiquer à des tiers à certaines occasions. Celles-là, nous gagnons tous et toutes en ne les confiant qu'à des tiers qui n'ont ni la capacité technique ni non plus intérêt à effectuer de la collecte massive.

3. LES NOUVEAUX PROTOCOLES ET LES NOUVELLES PLATEFORMES ET APPLICATIONS SOCIALES – DÉCENTRALISÉES OU DISTRIBUÉES – PROMUS PAR LE MILIEU DU LOGICIEL LIBRE POUR « RÉPARER » INTERNET ET SAUVEGARDER LA DÉMOCRATIE AU 21e SIÈCLE

Pour migrer nos données dans des écosystèmes numériques fait de services décentralisés ou distribués, pour profiter pleinement des avantages du chiffrement de bout en bout, il nous faut produire et adopter de nouveaux logiciels reposant sur de nouvelles architectures.

Heureusement, depuis 2013, de nombreuses actions ont été entreprises dans les milieux des hackeurs et des libristes afin de « réparer » Internet par la redécentralisation de ses applications et la promotion du chiffrement de bout en bout des données des internautes.

Dans ces milieux, on est peu optimiste quant à la capacité des sociétés démocratiques actuelles de produire un droit capable de suivre les

changements sociaux rapides qu'entraîne l'informatique, surtout dans un contexte où il faut se battre à la fois contre l'idéologie « sécuritaire » de l'après 11 septembre 2001 et les plus puissantes multinationales que le monde ait jamais connues. Beaucoup se sentent plus utiles et compétents sur le terrain de la réponse technologique, qui doit nécessairement accompagner la réponse politique à la centralisation et à la surveillance de masse qu'elle facilite.

Parmi les nombreuses initiatives visant à « réparer » Internet qui sont apparues récemment, on peut citer en exemple nul autre que l'inventeur du Web, le britannique Tim Berners Lee, qui se consacre depuis 2015 à un projet nommé Solid[12]. Avec ce projet, il espère susciter un mouvement de redécentralisation de son invention, qu'il considère à la dérive depuis que les GAFAM ont réussi à opérer une terrifiante centralisation organisationnelle et même sociale des usages d'Internet sur les fondations pourtant techniquement décentralisées d'Internet et du Web.

Comme tous les autres projets de redécentralisation du genre, Solid mise par principe sur les logiciels libres, les normes et les standards libres et ouverts, la démocratisation de la cryptographie forte et ultimement la reprise de contrôle par les citoyens et citoyennes de leurs données et de leurs identités numériques.

Ces initiatives sont pour l'heure à la marge, mais elles sont peut-être promises à un bel avenir dans le contexte de l'entrée en vigueur en Europe du Règlement général sur la protection des données (RGPD), qui crée un droit à la portabilité des données. Rappelons que l'objectif visé par ce droit est double : 1) aider les citoyens à reprendre le contrôle sur leurs données et 2) stimuler la concurrence entre les responsables de traitement de données[13].

3.1. Quelques exemples de protocoles et d'applications sociales libres et décentralisés

Parmi les protocoles les plus dignes d'intérêt pour l'émergence d'un nouvel environnement d'applications sociales libres et décentralisées, il y

12. https://solid.mit.edu

13. Par le *Consumer Privacy Act* voté à l'été 2018, la Californie s'est engagée dans le même sens que l'Europe sur la question du contrôle par les citoyens des usages de leurs données. Le mouvement s'étendra-t-il au Québec et au Canada ?

a *ActivityPub*, qui est une norme ouverte recommandée par le World Wide Web Consortium (W3C) depuis janvier 2018. Ce protocole normalisé, qui repose sur le format de données ActivityStream 2.0, décrit comment communiquer des messages qui peuvent circuler entre plusieurs sites web. Qu'est-ce que ça veut dire ? Un exemple permet de l'illustrer simplement. Avec YouTube (Google), une plateforme centralisée, il n'est pas plus possible de laisser des commentaires à propos d'une capsule vidéo qu'avec un compte YouTube. En implémentant le protocole ActivityPub du côté de YouTube, il deviendrait possible de commenter ladite capsule vidéo à partir d'autres applications sociales compatibles avec ActivityPub. Disparaîtrait donc la nécessité d'ouvrir un compte YouTube uniquement pour interagir avec une personne qui nous transmet une capsule vidéo de cette plateforme. Adopté de manière généralisé, ce protocole permettrait de décentraliser non seulement le fait de commenter, mais tout un ensemble d'« activités » que les internautes réalisent en ligne dans leurs interactions sociales : publier, mettre à jour, supprimer, accepter, refuser, suivre, ne plus suivre, etc.

Au moment d'écrire ces lignes, tout un écosystème d'applications sociales libres et décentralisées se développe autour d'ActivityPub avec Mastodon, NextCloud, PeerTube, etc.

Décrivons brièvement quelques-unes de ces applications.

Mastodon se présente sans gêne comme un clone de Twitter, mais qui offre l'avantage d'être libre, décentralisé et capable de gérer des messages allant jusqu'à 500 caractères. Apparu très récemment en 2016-2017, il a connu un succès étonnant avec plus de deux millions d'utilisateurs et d'utilisatrices (2019), qui ont ouvert leur compte sur l'un ou l'autre des quelque 2 500 serveurs fédérés du réseau[14].

NextCloud est une imposante suite d'applications web dont les principales concernent le stockage et le partage de fichiers, de contacts, de calendriers et autres essentiels du travail collaboratif en ligne. Les différents « nuages » NextCloud (les serveurs), qui peuvent être personnels, communautaires, commerciaux, privés ou publics, peuvent fonctionner à l'intérieur d'une vaste fédération décentralisée. ActivityPub permet une grande interaction entre les utilisateurs et les utilisatrices de NextCloud sans les obliger tous et toutes à ouvrir un compte sur une même instance centralisée de NextCloud.

14. Voir les statistiques recueillies sur le site https://the-federation.info

PeerTube est une plateforme d'hébergement vidéo libre, décentralisée et fédérée, qui se présente comme une solution de remplacement à YouTube, Vimeo, Dailymotion, etc. En plus d'être compatible avec ActivityPub, elle intègre le pair-à-pair grâce au protocole WebTorrent, qui permet de décharger intelligemment les serveurs PeerTube lorsqu'un grand nombre d'internautes visionne simultanément le même fichier vidéo.

3.2. Quelques exemples de protocoles et d'applications sociales libres et distribués (pair-à-pair)

Si la liberté de choisir à quel serveur on se connecte offre des avantages certains, il faut quand même ultimement faire confiance à un tiers qui pourrait nous trahir ou être compromis pour une raison ou une autre, malgré toutes nos précautions. C'est là qu'interviennent les applications sociales libres et distribuées dans lesquelles la relation client-serveur hiérarchisée se transforme en relation pair-à-pair égalitaire.

Comme c'est le cas de l'univers qui prône la décentralisation via des fédérations de serveurs décentralisés, l'univers qui prône le pair-à-pair regorge de projets d'applications, de plateformes et de protocoles visant à nous déprendre collectivement et individuellement des GAFAM en s'attaquant aux problèmes de la centralisation.

À titre d'exemples, regardons une application sociale distribuée (Jami) et un protocole de stockage distribué (IPFS).

Jami (anciennement *Ring*) se présente comme une plateforme de communication universelle et libre conçue spécifiquement pour la protection de la vie privée et plus généralement, la protection des libertés de ses utilisateurs et de ses utilisatrices. Jami est d'un côté un réseau utilisant des tables de hachage distribuées (*distributed hash tables*) pour relier directement tous les pairs, de l'autre une plateforme sur laquelle est construite une suite extensible d'applications de communication permettant de faire des appels audio, des appels de visioconférence, d'envoyer et de recevoir des textos, de partager des photos et d'autres fichiers, etc. Dans ses fonctionnalités principales, Jami se compare donc à Skype, WhatsApp, Google Hangouts et autres applications du même genre. Notons au passage que cette plateforme est développée principalement à Montréal par l'entreprise québécoise Savoir-Faire Linux.

InterPlanetary File System (IPFS) est un protocole qui vise à distribuer sur le réseau le stockage et le partage des hypermédias du Web. Ultimement, IPFS se veut une solution de remplacement au protocole HTTP(S), dont il corrige certaines des limitations liées à son architecture client-serveur. Pour l'heure, il est possible de faire communiquer IPFS et HTTP(S) via des passerelles de communication et des extensions de navigateurs web comme Firefox ou Chrome.

4. QUELQUES BATAILLES QUI NOUS ATTENDENT POUR SORTIR LES APPLICATIONS SOCIALES LIBRES, ÉTHIQUES, DÉCENTRALISÉES ET SOLIDAIRES DES MARGES OÙ ELLES SONT AUJOURD'HUI EN 2020

Développer de nouvelles applications sociales sur la base de critères éthiques adaptés aux défis de notre époque – notamment en matière de protection des libertés et des droits au fondement de la démocratie – implique nécessairement le choix des quatre libertés du logiciel libre, de la décentralisation ou de la distribution et du chiffrement fort, préférablement de bout en bout chaque fois que c'est approprié. Heureusement, plein de projets prometteurs allant dans ce sens sont en cours de développement et certains d'entre eux ont même déjà atteint un bon niveau de maturité.

Cependant, développer des logiciels plus éthiques n'est pas tout : il faut encore les faire adopter par le plus grand nombre, dans un contexte difficile où les logiciels non libres et les services centralisés sont préinstallés et préconfigurés sur les appareils numériques vendus aux consommateurs et aux consommatrices. Cette situation dure depuis de nombreuses années : les logiciels non libres et les services centralisés gratuits sont la normalité culturelle pour des millions de personnes. Les défis sont grands et il nous faudra mener bataille sur les terrains politique, social, économique et culturel si nous voulons gagner.

La bataille de la *visibilité* des alternatives et de leur *adoption massive* mobilise déjà certains milieux. En effet, tandis que certaines personnes et certains organismes mettent leur énergie dans le développement de nouveaux logiciels, de nouveaux protocoles ou de nouvelles technologies, d'autres choisissent plutôt de faire dans ce qu'on pourrait appeler de la « médiation » entre les communautés de développeurs et de développeuses et le grand public des utilisateurs et des utilisatrices d'appareils numériques. Des formes de médiation sont absolument nécessaires à bon nombre de

projets de logiciels libres, ceux-ci ne disposant de rien qui s'apparente aux budgets de publicité et de mise en marché des grandes entreprises qui sont propriétaires des logiciels non libres les plus utilisés.

Si on parle en particulier de favoriser l'adoption par le grand public des logiciels libres utiles à la décentralisation et à la reprise de contrôle sur nos données, deux actions récentes sont sans conteste dignes de mention : la campagne d'éducation populaire « Dégooglisons Internet » menée par l'association française Framasoft de septembre 2014 à septembre 2017 et le lancement en octobre 2016 du collectif des CHATONS.

Quelques détails sur ces deux actions permettront de comprendre pourquoi elles suscitent l'enthousiasme.

4.1. DéGAFAMiser Internet… en adoptant des CHATONS locaux !

Framasoft était déjà un réseau d'éducation populaire assez connu en France et dans la francophonie[15] lorsque sa petite équipe s'est lancée dans la campagne « Dégooglisons Internet » en septembre 2014, en réponse aux révélations d'Edward Snowden de l'été 2013, mais plus encore aux pièges des services en ligne « gratuits » dans lesquels même Framasoft était tombé éventuellement en utilisant Gmail, Google Groups et Google Analytics[16]. La campagne, audacieuse dans son nom même, consistait en somme à éduquer le public sur les enjeux et les dangers de l'Internet centralisé des GAFAM et à lui proposer des solutions de remplacement concrètes sous la forme de services libres, éthiques, décentralisés et solidaires, prêts à l'emploi et prêts à être déployés ailleurs grâce à des tutoriels en français. C'est ainsi que sont apparus les Framapad (c'est-à-dire une instance d'Etherpad, pour remplacer Google Docs), Framatalk (Jitsi Meet, pour remplacer Skype), Framadrive (NextCloud, pour remplacer Dropbox), Framaforms (Drupal Webform, pour remplacer Google Forms), etc. La plupart des services en ligne de Framasoft étaient (et sont toujours) des instances de logiciels libres développés ailleurs qui attendaient justement qu'on les découvre et qu'on les déploie massivement. C'est donc en tant que médiateur entre les

15. Fondée en 2001, Framasoft s'est d'abord fait connaître par son annuaire de logiciels libres en langue française disponible sur Internet, la Framakey, une compilation de logiciels libres pour Windows distribuée sur clé USB, et la maison d'édition Framabook, qui publie des livres (manuels, essais, bande-dessinées, romans, etc.) sous licence libre.

16. Framasoft, « Dégooglisons Internet : notre (modeste) plan de libération du monde », *Framablog*, 7 octobre 2014. https ://framablog.org/2014/10/07/degooglisons-internet/

développeurs de logiciels libres et le grand public que Framasoft est intervenu : en montrant que des solutions de remplacement existaient, qu'elles étaient dans bien des cas plutôt simples d'utilisation, et qu'elles pouvaient être déployées sous forme de services libres moyennant quelques efforts : des efforts à la portée d'une simple association sans but lucratif comme Framasoft.

En cours de route de la campagne, vers 2016, les gens de Framasoft ont souhaité profiter du succès populaire et de l'attention médiatique relative dont ils faisaient l'objet non pas pour croître et multiplier toujours plus le nombre des utilisateurs et des utilisatrices de leurs Frama services, mais pour tenter de faire connaître l'ensemble des hébergeurs ou fournisseurs de services numériques déjà engagés sensiblement dans la même démarche qu'eux. C'est dans cet esprit qu'en octobre 2016 Framasoft initiait le collectif des CHATONS[17].

Les CHATONS (Collectif d'Hébergeurs Alternatifs, Transparents, Ouverts, Neutres et Solidaires) sont quelque 50 hébergeurs ou fournisseurs de services numériques qui ont signé un manifeste et qui adhèrent à une charte éthique. Les points communs de tous ces fournisseurs sont essentiellement le recours exclusif aux logiciels libres, la transparence concernant les mesures prises pour respecter les libertés et les droits des citoyens et des citoyennes et les efforts de communication et de pédagogie concernant les enjeux et les défis du numérique à l'ère des GAFAM.

Les modèles économiques reposant sur la collecte massive des données, le pistage et le profilage sont évidemment exclus, mais autrement il n'y a pas de restrictions à ce niveau. On trouve donc dans le répertoire des CHATONS des organismes sans but lucratif, des coopératives, des regroupements informels, des entreprises « traditionnelles » et même des personnes qui offrent simplement leurs services à leurs amis ou leurs parents.

17. Framasoft, « Naissance du collectif CHATONS », *Framablog*, 12 octobre 2016. https://framablog.org/2016/10/12/naissance-du-collectif-chatons/

4.2. Services FACiLes : amener les CHATONS au Québec

Malgré l'ambition internationale du collectif, les membres des CHATONS sont presque tous européens (et surtout français) encore aujourd'hui en 2020[18]. Il y a des exceptions, notamment au Québec, grâce aux efforts de FACiL et de Koumbit.

En septembre 2015, dans le cadre de la Semaine québécoise de l'informatique libre (SQiL), FACiL invitait le délégué général de Framasoft, Pierre-Yves Gosset, à faire connaître la campagne « Dégooglisons Internet » aux Québécois. Quelques mois plus tard, lors de son assemblée générale annuelle de mai 2016, FACiL votait comme objectif de son nouveau plan d'action triennal la réalisation de la phase 1 du projet Services FACiLes consistant à « offrir aux Québécois(es) une petite gamme de services libres, éthiques, décentralisés et solidaires, dans le cadre du Collectif d'Hébergeurs Alternatifs, Transparents, Ouverts, Neutres et Solidaires (CHATONS) sollicité par nos ami·e·s libristes de Framasoft. » FACiL annonçait publiquement sa volonté de rejoindre les CHATONS dès le lancement du collectif en octobre 2016. Koumbit a fait de même peu après.

Malgré le peu de ressources de FACiL, le projet a fait son petit bonhomme de chemin, principalement grâce au travail de bénévoles.

Le 30 avril 2018, Date FACiLe, une instance du planificateur de rendez-vous Framadate, devenait le premier Service FACiLe à entrer en production « bêta ».

Un an plus tard, en avril 2019, cinq Services FACiLes étaient en production « bêta ».

Au-delà de la simple adoption massive des logiciels libres : l'autofinancement populaire et pérenne est-il possible ?

Si l'expérience a montré que le téléchargement gratuit de certains logiciels libres de qualité a grandement aidé à les faire adopter massivement, ce n'est clairement pas le cas de tous les logiciels libres dignes d'intérêt ! Non seulement il existe un grand nombre de logiciels libres de qualité qui sont à la marge face à leurs équivalents non libres, mais en plus ceux qui par chance ont connu une grande adoption ne trouvent pas automatiquement

18. Soulignons qu'il y a d'autres projets comparables en dehors des CHATONS : RiseUp.net, allmende.io, disroot.org, austistici.org, xnet-x.net, plateform.coop et d'autres. Voir https://wiki.chatons.org/doku.php ? id=reseau_international

les sources de financement nécessaires à leur pérennité dans le temps. Des campagnes de promotion continue comme celles de Framasoft, de l'April, de FACiL et d'autres favorisent sans doute l'adoption des logiciels libres, mais même si les succès devaient être retentissants, il resterait encore une grande bataille à remporter : celle de la *viabilité économique*. C'est en partie à ce grand problème que s'attaque la plus récente campagne de Framasoft nommée « Contributopia » (sept. 2017 à sept. 2020), qui fait suite à « Dégooglisons Internet » [19].

Au moment d'écrire ces lignes, Framasoft a réussi à recueillir auprès du public le financement nécessaire au développement de deux nouveaux logiciels libres utiles à la décentralisation dans le cadre des activités de « Contributopia ». Le premier de ces logiciels libres est PeerTube[20], présenté sommairement plus haut dans l'article, et le second est Mobilizon[21], qui vient d'obtenir son financement et dont le développement débute à peine. Mobilizon se veut une solution de remplacement libre et décentralisée aux événements Facebook, à Meetup.com et autres plateformes centralisées comparables.

La question se pose : le financement participatif peut-il solutionner le problème de la viabilité économique des logiciels libres pour lesquels les modèles économiques marchands (ceux basés sur les services informatiques comme le soutien technique, la formation, le développement sur mesure, l'hébergement, le conseil, les produits dérivés, etc.) ne fonctionnent pas bien ou pas du tout ? La réponse à cette question n'est pas simple.

Si le financement participatif a déjà amplement montré qu'il pouvait servir au démarrage de projets commerciaux et non commerciaux de toutes sortes (Kickstarter, Indigogo, etc.), il semble y avoir une volonté chez les libristes de l'adapter à leur éthique (la plateforme doit être un logiciel libre) et à des projets qui nécessitent un financement soutenu et récurrent : c'est

19. Framasoft, « Contributopia : dégoogliser ne suffit pas », *Framablog*, 9 octobre 2017. https://framablog.org/2017/10/09/contributopia-degoogliser-ne-suffit-pas/ et https://contributopia.org/fr/
20. Framasoft, « PeerTube bêta : une graine d'alternative à YouTube vient d'éclore », *Framablog*, 21 mars 2018. https://framablog.org/2018/03/21/peertube-beta-une-graine-dalternative-a-youtube-vient-declore/
21. Framasoft, « Mobilizon : finançons un outil pour sortir nos événements de Facebook ! », *Framablog*, 14 mai 2019. https://framablog.org/2019/05/14/mobilizon-financons-un-outil-pour-sortir-nos-evenements-de-facebook/

le cas notamment de Liberapay[22] et d'Open Collective[23], qui sont des équivalents libres de Patreon ou de Tipeee.

En 2019, l'autofinancement populaire des communautés de logiciels libres, qui repose sur les dons récurrents (parfois déductibles d'impôt), est rarement en mesure de transformer le travail bénévole en travail bien rémunéré. Pour beaucoup de communautés et de projets, on observe plutôt qu'il est une source de revenus parmi d'autres, dans le cadre de modèles économiques *hybrides* (marchands et non marchands). C'est dans le cadre de ces modèles économiques, parfois inédits, que les communautés de logiciels libres explorent les territoires de l'innovation sociale, où se mêlent les visés parfois antagonistes de personnes et d'organisations qui se réclament tantôt de l'économie libérale, tantôt de l'économie sociale et solidaire, tantôt du mouvement des communs, etc.

4.3. Vers des appareils numériques et des services en ligne certifiés éthiques ?

Rendre les logiciels libres populaires et viables économiquement est nécessaire aux efforts visant la redécentralisation d'Internet, mais est-ce suffisant pour opérer une vaste transition numérique, qui laisserait derrière les modèles économiques des GAFAM ? Rappelons que les logiciels de ces entreprises, des systèmes d'exploitation aux applications du quotidien, sont préinstallés sur les appareils numériques vendus dans toutes les surfaces commerciales, grandes ou petites. Ces appareils sont par ailleurs préconfigurés pour favoriser l'adoption des services centralisés de ces mêmes entreprises, qui accompagnent donc le public dans tous ses « choix » de consommateurs : lorsqu'il magasine ses appareils, ses applications, ses contenus culturels (musique, films, livres, jeux, etc.), et ses abonnements aux divers services qui s'occuperont de stocker et de manipuler ses données, même les plus privées.

N'y a-t-il pas moyen de mettre en marché des appareils numériques différents, sur lesquels seraient préinstallés des logiciels libres, à partir desquels le public se verrait offrir clé en main des services décentralisés, voire distribués ? Oui, c'est possible, et certaines initiatives qui vont dans ce sens méritent d'être mentionnées. Je me limiterai à trois d'entre elles

22. https://liberapay.com/
23. https://opencollective.com/

pour illustrer la tendance : la certification Respect Your Freedom de la Free Software Foundation, la ligne de produits Librem de l'entreprise américaine à vocation sociale Purism, et la Freedom Box de la Freedom Box Foundation.

Respect Your Freedom (RYF) est le nom d'un programme de certification éthique de la Free Software Foundation (FSF) qui s'applique aux appareils numériques de toutes sortes, des ordinateurs complets, en passant par les périphériques (imprimantes, webcam, etc.) et même les composants (cartes de son, cartes graphiques, cartes réseaux, etc.). Les critères définis par la FSF sont à la fois sévères et en même temps insuffisants en ce qui concerne les spécifications matérielles. Ils sont sévères en ce qu'ils exigent que le produit certifié non seulement ne soit accompagné que de logiciels libres et rien d'autres, mais qu'en plus il n'encourage d'aucune manière l'utilisation de logiciels non libres. Cette sévérité ne s'étend cependant pas au matériel informatique : un produit certifié RYF n'exige pas que 100 % du microcode d'un appareil soit libre ou que tous les plans et toutes les spécifications permettant la fabrication du matériel soient libres.

Comme on peut le constater en consultant la liste des produits certifiés RYF en 2019, il reste énormément de travail à faire. Il n'y a en effet pour l'heure que deux « grands » fournisseurs de produits certifiés pour la planète entière : ThinkPenguin (États-Unis) et Technoethical (Roumanie). Heureusement, il existe un nombre beaucoup plus grand de produits qui ne sont pas certifiés, mais qui pourraient sans doute l'être à l'avenir avec quelques efforts supplémentaires.

Librem est une ligne de produits de Purism, une entreprise fondée aux États-Unis en 2014. Les premiers Librem étaient des portables avec écran de 15 pouces et ensuite 13 pouces (respectivement Libre 15 et Librem 13) financés par des campagnes participatives. En 2017, l'entreprise a été réincorporée pour être reconnue légalement comme une entité à vocation sociale sous les lois américaines. La vocation de l'entreprise est de fabriquer des appareils numériques pour le grand public en considérant comme prioritaire la sécurité, la vie privée et les libertés des utilisateurs et des utilisatrices. Le Librem 5, un téléphone avec écran de 5 pouces, a levé son financement en ligne en 2017.

Freedom Box est une distribution (dérivée de Debian) qui permet de monter facilement un petit serveur personnel installable sur un ordinateur monocarte comme le Raspeberry Pi. Le serveur personnel permet d'installer en un seul clic tout un ensemble de logiciels libres permettant de décentraliser Internet par l'autohébergement. Quelques entreprises

(notamment Olimex) commercialisent des appareils préconfigurés avec Freedom Box.

Il existe d'autres distributions du même genre, notamment YunoHost et Cozy Cloud.

4.4. Le front politique : les réformes qu'il nous faut obtenir

La migration de nos données dans des environnements logiciels plus dignes de notre confiance et moins toxiques pour la démocratie s'annonce longue et complexe et il nous sera impossible d'arriver à une situation satisfaisante et idéalement stable en négligeant de mener bataille sur le front *politique*.

L'action politique (au sens classique du parlement et de l'action publique des gouvernements) pourrait facilement faire l'objet d'un article séparé vu son importance. Il en va de même pour l'action juridique, dont je ne parlerai pas. Je vais me limiter ici à quelques grandes généralités en rapport direct avec la transition vers un monde numérique meilleur sur un horizon « moyen » (peut-être 10-15 ans).

Pour mettre fin à la surveillance de masse réalisée à des fins commerciales, il faut globalement cesser d'utiliser des logiciels non libres, notamment ceux déployés en mode hypercentralisé par les GAFAM dans leur offre de services en ligne « gratuits ». Pour appuyer la transition des citoyens et des citoyennes hors des écosystèmes numériques des grandes multinationales américaines, plusieurs réformes peuvent être réalisées. Aux côtés de réformes du même type que celles qui ont donné le Règlement général sur la protection des données (RGPD) en Europe (a), on peut penser à des actions visant à mettre fin à la vente forcée des logiciels non libres avec les appareils numériques (b), donner la priorité aux logiciels libres et aux normes et standards libres et ouverts dans les secteurs public et parapublic (c), soutenir l'émergence d'un secteur de l'économie sociale et solidaire voué à la construction d'un monde numérique compatible avec les libertés et les droits fondamentaux, le gouvernement démocratique et l'environnement (d).

Pour ce qui est de la surveillance exercée par les États ou les villes, je ne vois pas d'autres solutions qu'une mobilisation importante de la part des citoyens et des citoyennes pour exiger – et protéger par de nouveaux droits – l'usage exclusif du logiciel libre, le recours au chiffrement fort

– notamment de bout en bout pour les communications – et la décentralisation ou la distribution pour le traitement et le stockage de nos données, qui doivent rester sous notre contrôle (e) et la réforme en profondeur du fonctionnement et des capacités d'agir des services de renseignement à tous les ordres du gouvernement (f).

CONCLUSION

Comme j'ai tenté de le montrer dans cet article, il n'y a que dans le milieu du logiciel libre qu'existe véritablement le souci de protéger les utilisateurs et les utilisatrices d'appareils numériques à la fois contre les filous qui cherchent à exploiter les failles de sécurité des systèmes numériques et aussi contre les abus et les erreurs des personnes et des entreprises qui conçoivent les logiciels qui traitent nos données.

On évite mieux les abus et on corrige mieux et plus rapidement les erreurs dans le cadre éthique des quatre libertés que les licences de logiciel libre reconnaissent aux utilisateurs et aux utilisatrices d'appareils numériques. Ces libertés peuvent non seulement être utilisées pour se protéger contre des attaques envers nos droits fondamentaux et nos intérêts, mais aussi pour contre-attaquer en quelque sorte en développant et en utilisant des logiciels libres, notamment des applications sociales, qui permettent de redécentraliser Internet (au niveau des infrastructures, des équipements, des données, des logiciels et de l'expertise) et ainsi d'échapper aux conséquences néfastes de la centralisation.

Il nous faudra recourir à toute la solidarité humaine dont nous sommes capables, à toutes les échelles d'action (locale, régionale, nationale et internationale), pour favoriser autant la *visibilité*, l'*adoption massive* que la *viabilité économique* à long terme des logiciels libres les plus pertinents pour le grand public dans ses usages quotidiens d'Internet.

Pour gagner, nous devrons relever de grands défis en menant bataille sur les terrains politique, social, économique et culturel.

BIBLIOGRAPHIE

Stallman, Richard, Williams, Sam, Masutti, Christophe, « Richard Stallman et la révolution du logiciel libre. Une biographie autorisée », Paris, Eyrolles, 2010, 324 p. https://framabook.org/richard-stallman-et-la-revolution-du-logiciel-libre-2/

Stallman, Richard, "Free Software, Free Society : Selected Essays of Richard M. Stallman", 3e édition, Boston, GNU Press, 2015, 293 p. (première édition : 2002). https://shop.fsf.org/books-docs/free-software-free-society-selected-essays-richard-m-stallman-3rd-edition

Lessig, Lawrence, « Code : Version 2.0 », New York, Basic Books, 2006, 410 p. (première édition : 1999) http://codev2.cc

4

ACCESS CONTROL IN CYBERSECURITY AND SOCIAL MEDIA

Nadine Kashmar

Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski, Rimouski, Québec
nadine.kashmar@uqar.ca

Mehdi Adda

Département de Mathématiques, Informatique et Génie, Université du Québec à Rimouski, Rimouski, Québec
mehdi_adda@uqar.ca

Mirna Atieh

Business Computer Department, Faculty of Economic Sciences and Administration,
Lebanese University, Hadat, Lebanon, matieh@ul.edu.lb

Hussein Ibrahim

Institut Technologique de Maintenance Industrielle (ITMI), Sept-Îles, Québec
hussein.ibrahim@itmi.ca

Nadine Kashmar is a PhD student at Université du Québec à Rimouski (UQAR), Canada. Her research project pertains to developing a new generic and enhanced access control metamodel and using it in the field of Internet of Things as a use case, specifically in Industrial IoT (IIoT). She holds her master's degree in Science of Computer Engineering from Beirut Arab University (BAU), Lebanon in 2016. She also has work experience in the field of IT and teaching experience in Computer Science. Her research interests focus on access control, IoT, Industry 4.0, and data mining.

Mehdi Adda is a professor of Computer Science at Université du Québec à Rimouski (UQAR), Canada since June 2010. From August 2008 to May 2010, he was an invited professor at the same university. His principal research interests lie in the fields of knowledge and data engineering, IoT and security. Mehdi Adda obtained two PhD degrees in Computer Science from Université de Montréal, Canada and Université de Lille, France in 2008. He received two MSc. degrees in Computer Science from Joseph Fourier University in 2002 (Grenoble, France), and Université du Havre (Le Havre, France) in 2003 as well as an Engineering degree in Computer Science from University of Sciences and Technology Houari Boumediene (Algiers, Algeria) in 2001.

Mirna Atieh obtained her PhD in Informatics and Artificial Intelligence in February 2008 from the Institut national des Sciences Appliquées INSA de Rennes, France. She is currently an Assistant Professor and Researcher at the Lebanese University in Lebanon – Faculty of Economic Sciences and Business Administration – Department of Business Computer. Her main research interests are in the areas of Artificial Intelligence (AI), Networking and Telecommunication, and Internet of Things (IoT). She has multiple scientific collaborations with various universities in France and Canada. She published several papers in international conferences and journals.

Hussein Ibrahim received his PhD degree in Engineering from Université du Québec à Chicoutimi (UQAC), Canada. From August 2009 to September 2016, he worked as research manager at TechnoCentre éolien in Gaspé. He has been working as research manager at Cégep de Sept-Îles in northern Quebec since September 2016, and as general manager of Institut Technologique de Maintenance Industrielle (ITMI) in Sept-Îles since September 2018. His research interests focus on renewable energy sources integration, hybrid energy power systems, storage energy, heat and mass transfer, energy efficiency, Industry 4.0 and IoT.

ABSTRACT

Social media networks and their applications (e.g. Facebook, Twitter...) are a current phenomenon with a great impact on several aspects such as, personal, commercial, political, etc. This media is vulnerable to various forms of attacks and threats due to the heterogeneity of networks, diversity of applications and platforms, and the level of users' awareness and intentions. As network technologies and their various applications evolve, the way people interact with them changes. Thus, the main concern for all social media users is to protect their data from any type of illegal access. With network and application developments, the concept of controlling access evolves in various stages. It begins with the implementation of the principles of information security (confidentiality, authentication ...), then by finding various access control (AC) models to enforce security policy in this field. For cybersecurity and social media, various methods are developed based on conventional AC models : Discretionary Access Control (DAC), Mandatory Access Control (MAC), Role Based Access Control (RBAC), Organization Based Access Control (OrBAC), and others.

In this chapter, we highlight the various types of cybercriminal attacks in social media networks. We then introduce the challenges faced for controlling users' access and the importance of the AC concept for cybersecurity and social media. We will also review the common AC models and the AC methods that are proposed to enhance privacy issues in social networks. Based on these methods, we will conclude our chapter by analyzing them to know their efficiency in such media and their adaptability for any future requirements.

1. INTRODUCTION

In social media, people create their own spaces to upload/share their private information (photos, videos, and audios) with family and friends using various forms of social networks, for example, Facebook, Twitter, LinkedIn, etc. Figure 1 shows the number of social network users worldwide from 2010 with a prediction up until 2020 (Rathore et al. 2017). Social networks allow users to extend their relationships and interact with strangers beyond family and friends. These networks have some interesting features (Rathore et al. 2017, Ali et al. 2018, Sayaf et al. 2014), as they:

- shorten the geographical distances between people worldwide;
- are used for entertainment, education, job searching, etc.;
- allow users to share their personal data with others in a relatively private manner;
- allow users to develop their social relationships by linking their profile with other users with similarities;
- users, companies or education sectors can create their own pages and post pertinent information in their timeline;
- used as an effective marketing strategy.

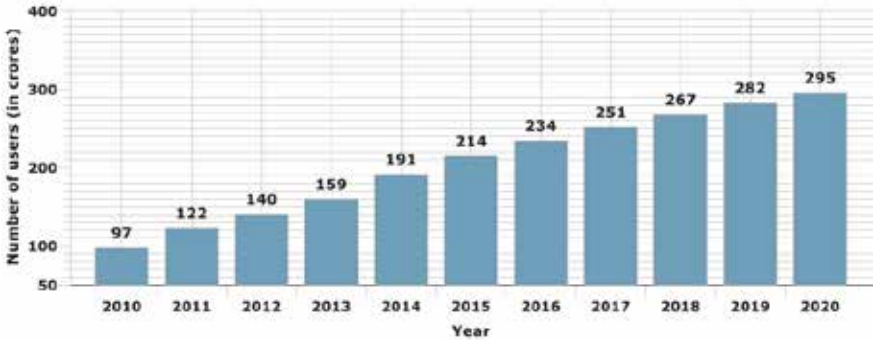


Figure 1 : Number of social network users worldwide from 2010 with prediction up until 2020

(Rathore et al. 2017)

Despite these features, we cannot ignore the fact that the increase in the number of social network users and the quantity of shared and uploaded information and multimedia data, result in a tremendous increase in security threats and vulnerabilities which affect users' confidentiality, authenticity and privacy (Rathore et al. 2017, Ali et al. 2018). Furthermore, a large number of users do not grasp the risks associated with what they post. Their lack of knowledge leads to an increase in cybercrimes. Social media security threats and vulnerabilities are generally divided into three categories (Rathore et al. 2017, Fire et al. 2014, Ali et al. 2018, Patsakis et al. 2015) :

a. *Traditional threats* : have been a problem since the widespread of the internet usage, this category includes many traditional attack techniques such as malware, phishing, clickjacking, etc.

b. *Social threats* : this category relates to the threats that intentionally target people of all ages (children, teenagers...) such as cyberbullying, cybergrooming, cyberstalking, risky behaviors, etc.

c. *Multimedia content threats* : when social media network users generate content and upload and share it online, they could be vulnerable to several risks from malicious behaviors such as multimedia content exposure, shared ownership, tagging, unauthorized data disclosure and more.

In the literature, a variety of solutions are proposed to deal with the above-mentioned threats. AC policies are considered as one of the possible solutions for privacy settings to measure and optimize security in social networks (Rathore et al. 2017). For this purpose, different AC methods to

enforce AC policies are developed to find secure communication environments and prevent any illegal access from attackers to users' information in social media networks (Sachan et al. 2011, Sayaf et al. 2014, Carminati et al. 2006). In this chapter, our concern is user security and privacy and for this purpose, we highlight the importance of AC mechanisms to mitigate security risks, then we explain some recent AC methods in this domain and how they are implemented to keep users' zones private and secure. This chapter is organized as follows. Social media network types and services are presented in section 2. The security threats and the possible vulnerabilities, also the importance of AC models for cybersecurity and social media are described in section 3. In section 4, the common AC models are summarized and the used AC methods in social media are also described. Finally, we conclude our chapter in section 5.

2. SOCIAL MEDIA NETWORK TYPES AND SERVICES

Current social media network services are web based. There exists numerous social media sites and applications with various purposes, services, and types which are developed over the years to include different types and categories. Table 1 summarizes these types of networks and provides some examples with a description for each type.

Table 1 : Social Media Network Types

Type	Examples	Description
Social/Relationship networks	Facebook, Twitter, Whatsapp, LinkedIn, Google+	People can connect and expand their relationships.
Media sharing networks	YouTube, Instagram, Snapchat, Facebook Live, Whatsapp, Flickr	Users can share photos, videos, and other media types.
Shopping networks	Wish, Gifteehub, AliExpress	Users can shop online, send gifts, share great finds, follow brands, etc.
Discussion forums	Reddit, Quora	Based on the posted subject, people can share news and ideas.

Type	Examples	Description
Bookmarking networks	Pinterest, Flipboard	Users can collect content they find interesting from the Internet, save and organize it, so it may be consulted at a later date (e.g. recipes, decorating ideas, etc.).
Interest based networks	Goodreads, Soundcloud, Houzz	Allow users to connect with other users with similar interests and hobbies.
Sharing economy networks	Airbnb, Uber, Rover, Taskrabbit	These types of sites allow people to advertise, find, share, buy, sell and trade goods and services.
Consumer review networks	TripAdvisor, Booking.com, Expedia, Local.com	These types of sites allow users to find, review and share information on products, services, hotels, restaurants, etc.
Blogging networks	Wordpress, BigCommerce, Wix, Medium, Ghost	Users can build their websites and publish content online, discover and comment.
Anonymous social networks	After School, Anomo, ASKfm, NoName	These types of applications allow users to post anonymous content and share on a private message board, their feelings, accolades, to gossip or snoop.

Social media network services are web based and need an internet connection. Users can use web sites via various devices (computers, smart phones…) with different platforms and applications. Some sites and applications combine more than one type, for example, Facebook is a social network site where users can share media (photos and videos). Social media networks allow people to expand their relationships (personal, business, education…), shop for various items, share information, advertise products or services, and express their feelings or to gossip. As shown in Table 1, social network types allow users of all ages to address various concerns and interests. Furthermore, it is well known that the internet world with all its related services, is vulnerable to numerous types of attacks. Hence, what

are the social media security threats and vulnerabilities? Which procedures and methods are required to preserve user security and privacy? A media with several types of sites and services and a large number of users of all ages, of different cultures and intentions, opens wide the doors for such inquiries. Section 3 summarizes social media security threats, vulnerabilities and cyberattacks.

3. SOCIAL MEDIA NETWORKS: USER SECURITY AND PRIVACY

Social networks let users communicate, share posts, interests, music, recommend books, movies and so on. The most serious concern for users in all aspects of their lives (e.g. personal, professional, entertainment...) is how to keep their zone secure and private. Hence, what are the possible risks when users post their private information on public networks? Before explaining the privacy requirements for social networks, it is important to present the security threats and vulnerabilities in this area.

3.1. Security Threats and Vulnerabilities

Social media users must realize that even when they use high security settings or websites, and even when they only select/accept their known friends, they are unintentionally leaking their information. Likewise, most people are unaware that the personal information they share/upload could lead to attacks against them or their friends. Hence, they must be aware and know the following:

- Once users post their information on a social network, there is no guarantee that this information is still private. As the quantity of information being posted increases, so does its vulnerability.
- It is more likely that, the more information is shared, an attacker or intruder could impersonate users and mislead their friends with actions such as download malware, share personal information, etc.
- On some social networks, information is publicly visible by default. Consequently, users should be mindful to change their privacy settings to *private* before posting information. Also, users must be aware that some other social networks change their privacy policy without their approval and that some of their information which was posted as private, may become public.

- Moreover, users should not expose themselves to several types of privacy and security issues. For example (Rathore et al. 2017, Deliri et al. 2015) :
  - it is preferable not to share a large amount of personal data on social networks ;
  - users should not post their location (home, work…) or children's location (school, summer camping…) ;
  - they should not provide their telephone number and credit card details, for example, for game applications ;
  - users should not accept friend requests from unknown people ;
  - most users do not read the privacy policy and terms of service for a social network before they create their accounts. They are therefore unaware of the policy nor of its updates.

Consequently, this could lead to various cybercriminal attacks. As mentioned earlier, security threats and vulnerabilities are generally divided into three categories : traditional threats, social threats, and multimedia content threats. Table 2 summarizes the various traditional attacks, such as spamming, phishing, etc. Through these types of attacks, an attacker tries to obtain personal information (password, bank account details …) for a user to commit to some critical attacks, for example, identity theft (Rathore et al. 2017, Ali et al. 2018, Fire et al. 2014, Zhang et al. 2018, Delerue et al. 2012, Deliri et al. 2015).

Table 2 : Summary of traditional security threats and vulnerabilities in social media

Threat	Description
Malware	A malicious software, consists of Trojan horses, viruses, or worms, used by attackers to assault users by sending injected scripts to the legitimate user and when clicking a malicious URL, a malware may be installed on devices and attempt to steal personal information from the victim.
Phishing	Attackers pretend to be a legitimate entity that the victim trusts, using fake websites and emails to expose a user's sensitive private information.
Spamming	Attackers send spam or junk data in bulk to internet users which causes network congestion.

Threat	Description
Clickjacking	A malicious mechanism used to make users click on something that is different from what they meant to click. For example, an attacker can manipulate users to post spam posts on their Facebook timeline, or use users' computer hardware (e.g. camera) to record their activities.
De-anonymization	Some users preserve their anonymity and privacy by using a false name. An attacker can find their identity by linking the information that the user has disclosed on a social network. This strategy is based on data-mining techniques where the attacker uses tracking methods, such as tracking cookies, or user group membership to expose the true identity of the user.
Identity/profile clone	An attacker can clone an existing user's profile in the same social media site or in a different one. The attacker then sends friend requests to the user's contacts and creates a trusting link with the real user's friends and collects sensitive information to carry out several types of scams (e.g. cyberbullying).
Inference attacks	An attacker can deduce a user's private information by exploiting other information that has been published about him on social media, such as data from the user's friends list. Then, this information is carried out using data mining techniques and can be used by the attacker to obtain internal secret information belonging to organizations.
Information leakage	Plenty of sensitive information can be inferred with great accuracy from material users share or post, or through data shared with other mutual users.
Sybil attacks & fake profiles	Attackers can create many fake identities, and by manipulating them, they can outvote the legitimate users. For example, they can boost the reputation and popularity of a user by voting him as the "best" over other legal users. They also have the capacity to corrupt information.

(Rathore et al. 2017, Fire et al. 2014, Ali et al. 2018, Zhang et al. 2018, Delerue et al. 2012, Deliri et al. 2015)

Table 3 summarizes the social security threats and vulnerabilities which occur in this media. Attackers can badly use the social relationship feature

of social networks. This feature enables attackers to interact with different types of users and in different ways. For example, they can entice teenagers by conveying sympathy, love, or care, or by offering money or gifts. Other behaviors might include espionage, blackmail, or sharing pornographic videos and images (Fire et al. 2014, Rathore et al. 2017, Deliri et al. 2015).

Table 3 : Summary of social security threats and vulnerabilities in social media

Threat	Description
Cyberbullying and cybergrooming	Cyberbullying is an intentional and iterative attempt by someone online harassing or harming. Cybergrooming is when a child or a teenager is humiliated and targeted with a malicious purpose by another teenager or child via the Internet. Cyberbullying and Cybergrooming are quite dangerous as it has led teenagers to extreme acts of violence, such as committing suicide.
Cyberstalking	Some users reveal their profile's personal information (e.g. phone number, home address, location…). This information can be tapped by malicious users for cyberstalking. For example, attackers can blackmail their victims through telephone calls or by sending instant messages using a social network site.
Risky behaviors	This may occur to teenagers or children through interactions with strangers in chat rooms, direct online communication, or giving private information and photos to an attacker. These behaviors can cause massive concerns regarding children or teenagers' safety.
Corporate espionage	Attackers use espionage techniques for commercial or financial purposes. Such techniques can be used by a competitor posing as a worker in the target company with the intent of spying on confidential information or hacking computers.

(Rathore et al. 2017, Fire et al. 2014, Deliri et al. 2015)

Table 4 presents the different multimedia security threats and vulnerabilities content in social media. Social network is another meaning for sharing data (photos, videos, interests, and so on). Multimedia data, especially high-resolution videos and images make it easier for estimating the location, geotagging, face recognition and more via multimedia retrieval techniques. Hence, the shared multimedia data can be illegally used by an

intruder to detect the user's location to find out if they are away from his home with the intention of theft (Fire et al. 2014, Rathore et al. 2017, Cutillo et al. 2010, Patsakis et al. 2015)

Table 4 : Summary of Multimedia Content Security Threats and Vulnerabilities in Social Media

Threat	Description
Multimedia content exposure	Posting multimedia data (videos, images, locations...) by users can expose them to various types of attacks, as they are disclosing an enormous amount of sensitive information. For example, intruders may follow the continuous posts of users' locations and when they are not home, this leaves the door open for intruders. Also, sharing a photo or video may violate other users' privacy if it is posted without their permission.
Metadata	Metadata provides information about other data. Multimedia content is considered as metadata since it contains a huge amount of other data, e.g. users' location tags, profession, family and more. Some of this metadata may be valuable to attackers when it is revealed.
Video conference	Most social network sites support video conferencing features (e.g. Facebook) which provide more interaction between users. An intruder may restrict the broadcasting of a video stream through vulnerabilities in the underlying communication architecture, or they can access the webcam of a user by using a malware program.
Tagging	Tagging is a feature within shared multimedia data made to increase interactions between users and to facilitate search capabilities. This feature could increase privacy risks for tagged users. Some users do not like to upload pictures of themselves on social media, as this feature can represent a violation to their privacy.
Hijacking	An attacker could gain control over someone and hijack his profile if the user has a weak password. It is preferable to use strong passwords in social media accounts and to change them frequently.
Shared ownership	Multimedia content (photos or videos) may correlate to many users, e.g. two people may take a photo at an event and one person can upload this photo with their preferred privacy settings without the permission of the other.

Threat	Description
Steganography	This is a tactic for concealing data within other media data. However, a malicious user can share malicious data by concealing it within multimedia data. For example, a picture with concealed malicious messages might be shared by a malicious user and a user may download it without knowing what it contains. This type of behavior is risky for the reputation of social network sites.
Manipulation of multimedia content	Malicious users can distort shared multimedia data by using available tools. For example, they can manipulate pictures to cause harm to others or to ridicule them.

(Rathore et al. 2017, Fire et al. 2014, Patsakis et al. 2015, Cutillo et al. 2010)

After exploring and explaining the aforementioned social media services and attacks, we can recognize that social media networks are the best environments for attackers to commit cybercrimes. In this context, various researches are conducted to resolve these threats and find the best ways to mitigate or prevent them, such as spam detection (Miller et al. 2014), phishing detection (Lee et al. 2013, Gupta et al. 2018), watermarking (bin Jeffry et al. 2017, Zigomitros et al. 2012), privacy settings (Ghazinour et al. 2016, Fiesler et al. 2017, Aldhafferi et al. 2013), authentication mechanisms (Joe et al. 2017, Ikhalia et al. 2013, Jain et al. 2015), steganalysis (Li et al. 2015, Taleby Ahvanooey et al. 2019) and other solutions. In the light of finding solutions for social media threats and vulnerabilities, we focus on privacy which is considered as one of the fundamental security objectives in social media environments (Cutillo et al. 2010, Zhang et al. 2010, Madejski et al. 2012, Sayaf et al. 2014). Privacy solutions include (Aldhafferi et al. 2013, Rathore et al. 2017, Patsakis et al. 2015) :

- Protective technologies such as strong authentication and AC mechanisms ;
- Users' awareness which addresses the issue of educating users about new technologies and explaining for them the possible risks of misusing social media networks.

In this chapter, our interest is to present and analyze the used AC methods in social media networks, to find their effectiveness and how they could prevent or mitigate security threats.

3.2. The Importance of Access Control Methods for Cybersecurity and Social Media Networks

In the era of social media, a huge amount of sensitive information can be easily collected, saved or deduced. Consequently, protecting users' privacy is the main objective for the services provided by social media platforms (Cutillo et al. 2010). The above-mentioned threats (section 3.1) clearly show that there are numerous security risks with the use of social media. To minimize social media security risks, various organizations have developed a formal policy to guide users on how to use social media sites for work-related activities (Delerue et al. 2012). A formal policy is the definition of guidelines, rules or regulations for a social network site (or an organization) to determine what is an acceptable or unacceptable use of social media, what information users can or cannot share, and the consequences for not following the defined policy. Guidelines examples for an organization can be as follows (Delerue et al. 2012, Cutillo et al. 2010):

- users should use strong passwords and update them regularly;
- users should not use the same password for a social media site as the one they use for their company;
- it is not allowed for users to share their organization's information or news on social networks.

Moreover, the defined social media security policies need to be effectively implemented for social media networks. Nevertheless, some users do not comprehend or ignore the privacy policy set by social network sites due to their level of understanding or to its complexity. Thus, they are unaware of the security risks that could occur when posting information (photos, videos...). In social media, AC mechanisms allow users to define their settings of privacy via control functions such as:

- the visibility of their own information;
- allow/deny others to write on their walls;
- determining the privacy of the shared contents (public, only friends, or user-defined group);
- Share posts with friends of friends, and many other privacy settings.

Various researches mention that (Deliri et al. 2015, Cutillo et al. 2010, Aldhafferi et al. 2013), users' authentication and AC functions must be powerful so that cybercrimes from cybercriminals, hackers, or spammers can be reduced as much as possible. For this purpose, different AC models

and mechanisms are developed to enforce privacy policy in social media networks. An AC framework lets users set their privacy preferences and allows application developers to create a customized plan based on users' preferences. In the following section, we first introduce the common AC models, then we review the AC models that are developed for social media networks.

4. THE ACCESS CONTROL MODELS

An access control model is a formalization for policies which are defined, by an organization for instance, based on a set of principles or guidelines for its system to control and authorize access to data. The common AC models implemented to prevent illegal disclosure of sensitive data and to protect data integrity are the following : Discretionary Access Control (DAC) (Hu et al. 2017, Ennahbaoui et al. 2013, Kashmar, Adda, and Atieh 2019), Mandatory Access Control (MAC) (Hu et al. 2017, Ennahbaoui et al. 2013, Ausanka-Crues 2001, Kashmar, Adda, and Atieh 2019), Role Based Access Control (RBAC) (Hu et al. 2017, Ennahbaoui et al. 2013, Kayem et al. 2010, Sandhu et al. 2000, Kashmar, Adda, and Atieh 2019), Organization Based Access Control (OrBAC) (Kashmar, Adda, and Atieh 2019, Ennahbaoui et al. 2013), and Attribute Based Access Control (ABAC) (Hu et al. 2017, Kashmar, Adda, and Atieh 2019, Kayem et al. 2010, Sandhu et al. 2000). Each model has its particular features and methods for making AC decisions and policy enforcement. Based on these models and due to various information technology concerns and needs in different fields, many other AC methods are extended and developed using features of two or more AC models (Kashmar, Adda et al. 2020). In the following sections, the common AC models and their features are summarized and some recent state-of-the-art AC methods for social media networks are described.

4.1. The Common Access Control Models

Each organization has its rules, guidelines and regulations which are defined as policies to control how users can access its logical and physical assets. An AC policy defines constraints on whether a user's access request to an object should be allowed or denied. Several AC methods are implemented at different information technology (IT) infrastructure levels, they are used in operating systems, databases, networks, etc. (Kashmar,

Adda, and Atieh 2019). The objectives of AC models can be summarized as follows:

- protect files and directories for organizations and information for all types of users;
- Regulate access to database objects and fields to protect application information such as payroll processing, e-health. etc.;
- Minimize the risk of unauthorized access to assets, thus minimize the risk to the business or organization.

Access right means that a subject is allowed or denied performing an operation on an object (Hu et al. 2017). AC policies might have the following form:

Allow managers to… and…

Knowing that… if… and/or…

Except…when…

Some AC policy examples can be written as follows:

- allow users A and B to read/write from/into file F for user C;
- allow technicians to read and follow the technical report instructions for machine *M*, during their working hours, if it is signed and confirmed by their technical manager;
- prevent social media users to send a friend request for user *A* if they are not friends of a friend.

Consequently, the main objective of AC models is the enforcement of the defined AC policies. In general, AC methods are defined in terms of subjects (e.g. user or program), objects (e.g. file, table or class) and access rights.

4.1.1. Discretionary Access Control (DAC)

The Discretionary Access Control (DAC) model was first introduced by Lampson (in the 1960s), a member of a curriculum design team. The three major components of this model are a set of objects, a set of domains, and a matrix (Kashmar, Adda, and Atieh 2019). Graham and Denning then extended Lampson's work where the term *subject* was included instead of the domain. Thereafter, Harrison, Ruzzo and Ullman (HRU) extended

Graham-Denning's work to find a more flexible model with the ability to describe several AC approaches (Hu et al. 2017).

This AC model is a user-centric model where a file owner can control the permission of other users requiring access to his file. Users can control the access rights (read, write, ...) to their files with the need of a pre-specified set Matrix called Access Control Matrix (ACM). Table 5 shows how AC rights of subject(s) over object(s) are specified. The intersection of u2 and o2 means that u2 can read the object o2. This ACM can also be implemented in two other variations, the first matrix is Capability Lists (CLs), and the second is Access Control Lists (ACLs). In CLs a user's access rights to access objects are represented by rows, while in ACLs the access rights for various users to access an object are represented by columns.

Table 5 : Access Control Matrix (ACM)

Objects

Subjects		$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$
	$\mathbf{u_1}$	read, write		
	$\mathbf{u_2}$	Update	read	
	u_3			delete

This model is provided with operating systems to authenticate system administrators and users using passwords.

4.1.2. Mandatory Access Model (MAC)

The Mandatory Access Control (MAC) model was presented in the 1970s to include the use of a security kernel. In this model, users cannot define AC rights by themselves. MAC is based on the idea of security levels which are associated with each subject and object. These levels have hierarchical and nonhierarchical components (Ennahbaoui et al. 2013, Hu et al. 2017, Kashmar, Adda, and Atieh 2019) :

- the hierarchical components include *unclassified (U)*, *confidential (C)*, *secret (S)*, and *top-secret (TS)* types where TS ≥ S ≥ C ≥ U, to categorize subjects and objects into levels of trust and sensitivity. For subjects, a security level is called *clearance level* and for objects it is called *classification level* ;

- the nonhierarchical components represent a set of categories where two security properties are used as security labels to indicate security levels for classification of objects and clearance of subjects, which are *simple property* and **-property*.

There are two variants for MAC, Bell and LaPadula (BLP) and BIBA (developed by Kenneth J. Biba). The first, *simple property* indicates no read up and *star property* indicates no write down. Hence, a subject is permitted to read an object if its clearance is ≥ than the object's classification, and to write if it is less than or equal (≤). The second, *simple property* indicates no read down and *star property* indicates no write up. Consequently, a subject is permitted to read an object if its clearance is ≤ than the object's classification, and to write if it is greater than or equal (≥) (Kashmar, Adda, and Atieh 2019, Ennahbaoui et al. 2013). Moreover, MAC standards are enforced by the operating system after they are defined by a system administrator. This model is proposed to overcome the limitations of the DAC model.

4.1.3. Role-Based Access Control (RBAC)

The Role Based Access Control (RBAC) model was proposed by David Ferraiolo and Richard Kuhn in 1992, it is developed as an alternative approach to MAC and DAC (Ennahbaoui et al. 2013). The RBAC approach is based on several entities which are users, roles, permissions, actions or operations, and objects. In this model, a role means a group of permissions to use object(s) and perform some action(s), and this role can be associated to several users. Also, users can be assigned to several roles based on their qualifications and responsibilities, such as accountants, directors, engineers, etc. (Hu et al. 2017). The RBAC model was implemented to facilitate the administration of the AC policy. It administers the access of a user to objects through roles for which the user is authorized to perform.

The RBAC can be applied in distributed systems because it is based on the concept of constraints and inherence. Role hierarchy determines which roles and permissions are available to subjects based on different inheritance mechanisms (Belokosztolszki 2004, Crampton 2003).

4.1.4. OrganizationBased Access Control (OrBAC)

The Organization Based Access Control (OrBAC) was presented in 2003 and proposed to solve some limitations in the previous models (DAC, MAC

and RBAC). Its aim is to find a more abstract control policy. Every organization (e.g. clinics, banks, hospitals…) is composed of a structured group of subjects having roles or entities. In OrBAC, seven entities are defined (Ennahbaoui et al. 2013) :

a. the abstract or organizational level composed of (1- Role, 2- Activity, and 3- View) ;
b. The concrete level constitutes (4- Subject, 5- Action, and 6- Object), and :
c. the seventh entity which is Context lies between the two levels to express dynamic rules for relations between entities, for example, Permission, Prohibition, Isprohibited, Recommendation, Ispermitted, Isobligatory, Isrecommended, Obligation between the elements of each level.

Thus, OrBAC exceeds the notion of granting permissions to subjects, it addresses the idea of prohibitions, obligations and recommendations. In such a way, a role may have a permission, prohibition or obligation to do some activity on some view given an associated context (Kashmar, Adda, and Atieh 2019).

4.1.5. Attribute-Based Access Control (ABAC)

The Attribute Based Access Control (ABAC) is the latest AC model development and its concepts have paralleled that of RBAC. ABAC has some advantages over RBAC, because of its ability to support dynamic attributes and its benefits in managing authorizations (Jin et al. 2012). ABAC has three types of attributes : object, subject and environmental attributes. This model allows or denies user requests based on some user attributes and on some other attributes of the object and environment. It is dynamic since it uses these attributes to determine access decisions (Hu et al. 2017), also AC permissions are evaluated at the time of the actual user's request. This offers a larger set of possible combinations of variables to reflect a larger set of possible rules to express policies (Crampton 2003). Hence, subjects are enabled to access a wider range of objects without specifying individual relationships between each subject and each object. This an ABAC advantage over RBAC. The Extensible AC Markup Language (XACML) and Next Generation AC (NGAC) are two standards that widely address the ABAC framework.

To summarize, the DAC model is proposed for the academic field, and the MAC model, for the military domain. RBAC includes features from DAC and MAC, and it is implemented to overcome some limitations of the previous models. OrBAC includes features from DAC, MAC and RBAC. It is proposed to find a more abstract control policy and to overcome the deficiencies in the previous models. RBAC has some limitations in supporting dynamic attributes such as the time of day. For this purpose, the ABAC model is proposed to support these attributes. Likewise, all the presented AC models still have some limitations (Kashmar, Adda, and Atieh 2019), this necessitates the need to find other AC methods with combined features from two or more AC models, or to integrate basic privacy requirements with the existing models. Consequently, different AC mechanisms for different computing environments are implemented. Upgrading or finding new AC mechanisms for the current challenging environments (social media networks, Internet of Things (IoT), cloud computing, etc.) is a critical requirement to follow the continuous upgrades and to mitigate security risks by preventing any illegal access (Kashmar, Adda, et al. 2019a, b).

4.2. Access Control Models in Social Networks

In the literature, various proposals exist to address the issue of AC in social networks. In general, as social media users are vulnerable to attackers, security concerns are expanded to include their privacy, financial transactions, their families and friends, cyber-theft threats and more. In addition to these concerns, it is unacceptable to allow a hacker impersonate users and trick their friends and families. Thus, social media services, the utilized technologies, security threats and vulnerabilities reflect the fact that social media and cybersecurity incorporate different security aspects, from social media sites to user behavior. In this context, several AC mechanisms are implemented with the intent to improve and preserve user privacy. AC methods are used in social media networks to enable users to control the propagation of their own data and protect their privacy against attacks (section 3).

In the subsequent sections, we review the AC methods pertaining to this domain. Some are implemented based on features of the common AC models, while others are implemented by considering the basic privacy requirements in social networks.

4.2.1. Access Control Methods based on Features of common AC models

In conjunction with the widespread use of social media network types and services (section 2), different AC mechanisms based on features of the common AC models are implemented. This section presents some of the proposed AC methods in this field.

Tie-RBAC : RBAC application to Social Networks

Tie-RBAC is the RBAC application to Social Network Analysis (SNA) (Tapiador et al. 2012). SNA provides a comprehensive body of concepts and methods for modeling social networks, it also provides the research sector with methods for social networks analysis (O'Malley et al. 2017, Tapiador et al. 2012). Hence, Tie-RBAC indicates *RBAC + SNA = Tie-RBAC*.

The objective of Tie-RBAC is for it to be implemented in a core for building social network sites. However, social entities or actors (a user, a group, a department, an organization…) are linked by social ties, where a tie is made up of two actors and a tie type. The first and the second actors are the sender and the receiver of the tie. Tie types between actors include emotional, formal or biological relationships, transfer of material resources, messages, conversations, affiliation to same organizations, etc., for example, a tie of friendship between actor A and actor B. Hence, a relationship in the network is the set of all ties of the same type between actors. This type of relationship is called reciprocal as in Facebook, some other relationships are non-reciprocal as in Instagram, where actor A may not follow actor B, but the opposite is true. Tie-RBAC is based on non-reciprocal ties, it is the RBAC application to social networks where :

- actors define their custom relationships (friend, partner, family etc.), which are equivalent to roles ;
- each actor assigns permissions to relationships, such as post to wall, read wall, etc. ;
- actors establish ties using these relationships where each tie is equivalent to the association of an actor to a role-relationship, as shown in Figure 2.

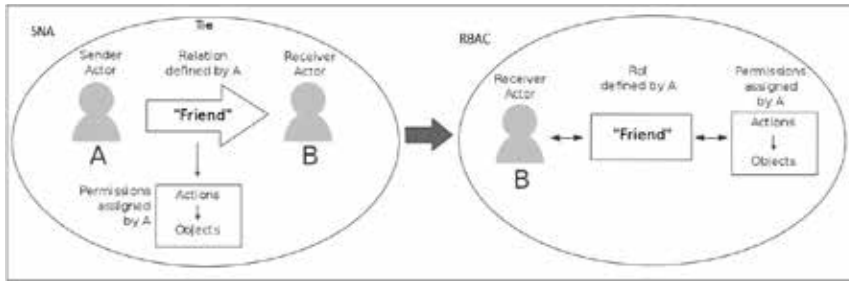


Figure 2: Tie-RBAC model: Equivalence between SNA's tie establishment and RBAC

(Tapiador et al. 2012)

When establishing the tie, the sender is the entity who grants privileges to objects and the receiver is the entity assigned to the role which gains permissions on the sender's objects. In the Tie-RBAC model, the relationship which is defined by the sender is the role. In Figure 2, actor A (sender) defines the relationship "friend" and allows "friend" to read the wall and post to it. A "friend" relationship is selected when establishing the tie with actor B (receiver). Thus, actor B is authorized to read the actor's A wall and post to it.

The purpose of this model is to provide social actors and web developers with a tool to build websites with social network features, and to define their own relationships which are adapted to their field of activity. Moreover, the AC enforcement in this model is the typical RBAC (Tapiador et al. 2012).

EASiER: Encryption-Based Access Control in Social Networks with Efficient Revocation

EASiER is an architecture that supports fine-grained AC policies and dynamic group membership by using Attribute-Based Encryption (ABE) (Jahid et al. 2011). The aim of EASiER is to shift AC policy enforcement from the social network provider to the user by means of encryption in order to mitigate the privacy risks in social networks. This case creates a key challenge in managing to support complex policies involved in social networks and dynamic groups. The key feature of this architecture, as mentioned by Jahid et al., is the possibility of removing access from a user without releasing new keys to other users or re-encrypting ciphertexts (CT). To handle this, a proxy is created to participate in the decryption process and enforce revocation restrictions. It also cannot provide access to users who are previously revoked. Figure 3 illustrates the EASiER architecture.

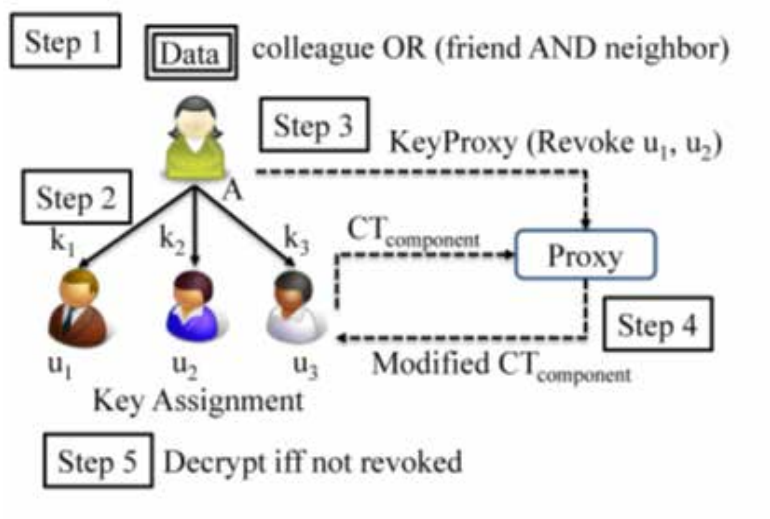


Figure. 3 : EASiER architecture

(Jahid et al. 2011)

The primary purpose of EASiER is to protect unintentional and intentional information leakage in social networks through ABE. EASiER allows users to :

- define relationships by assigning attributes and keys to each other ;
- create groups by assigning different attributes and keys to their social contacts ;
- encrypt different parts of data such as profile information, wall posts, etc. with attribute policies ;
- only contacts with keys having sufficient attributes that satisfy a policy can decrypt the data.

As shown in Figure 3, user or actor A can :

- define the attributes (friend, colleague, neighbor) and ;
- create keys k_1, k_2, and k_3 for the grouping of attributes of "colleague, friend, neighbor" and for "colleague, neighbor". Keys k_1, k_2, and k_3 are then assigned to u_1, u_2, and u_3. User A can also encrypt his/her data with the policy "colleague or (friend and neighbor)" ;
- user A may wish to end the relationship with u_1 and u_2 by revoking the corresponding keys which allow them to view A's data encrypted with any policy that their keys satisfactorily meet. Also,

user A may wish to revoke the attribute "neighbor" from k_3 which is assigned to u_3 and do a corresponding change in access control. The proxy of each user is assigned a secret proxy key with revocation information;

- it then uses its key to transform CT into a form with sufficient information and
- that an unrevoked user can mathematically combine with his secret key, then perform decryption where a revoked user cannot do so. The proxy key allows the disclosing of the components during a decryption for unrevoked users, whereas revoked users are unable to decrypt any data since they will not get assistance from the proxy. (Jahid et al. 2011).

The EASiER mechanism does not allow the proxy to decrypt data if it does not have the attribute keys. Additionally, a new proxy key is created each time a revocation is done, hence revoked users are prohibited from conspiring against each other or the proxy to get the data. However, only particular users who have the required set of attributes can decrypt the data (Jahid et al. 2011).

Organization Based Access Control Model for Social Network

An OrBAC extension is implemented by Belbergui et al. (2016) and adapted to the Facebook context. OrBAC is an AC model based on the organization, the first-order logic is used to define relations between entities and AC policy which is defined on two levels. The first, is the abstract level (role, activity, view), and the second, is the concrete level (subject, action, object). Policy levels are adapted to the context of Facebook as follows:

- friends are defined by role (friends, friends of friends, family, etc.)
- actions are classified by activities (display, publish, etc.) and
- account owners' data is organized by views (personal information, photos, etc.).

As stated by Belbergui et al. (2016), the process of modeling the Facebook AC policy using OrBAC model is simulated based on the inventory of roles (friends, family, etc.), of activities (create, consult, etc.), of views (personal information, etc.), and of access rights (permissions). Simulation of security policy with MotOrBAC simulator is also illustrated for:

- creating organizations (Facebook, U1, etc.);

- adding abstract entities (roles, activities, views) ;
- adding concrete entities (subjects, actions, objects) ;
- adding access rights
- simulation : detection of conflicts.

Facebook is defined as an organization, users are defined as a sub-organization of Facebook, and users' accounts (u_1, u_2...) are defined as a sub-organization of users. Roles are defined to be the usage by all users such as friends, families, studies, etc. Facebook users keep their photos, videos, etc. where other members/users are authorized/denied certain actions such as viewing pictures, writing on a wall, etc. These actions (open, view, read, search, share, etc.) can be structured into activities (create, consult, publish, block, accept, etc.). In other words, the entities *actions* define how subjects can access objects, and the structuring of these entities is called *activities*. Linking these entities is called a relation, for example, consider (org, a, a) means that the organization org considers action a, as part of activity a. Other relations exist between views and organization objects to facilitate the management of the security policy.

Based on the aforementioned concepts, Facebook-User policy and User-User policy are defined. Table 6 shows some examples for the defined policies.

TABLE 6 : Examples of Facebook-User and User-User defined AC policies

Facebook-User Policy	**User-User Policy**
Permission (Facebook, users, create, account)	Permission (u_1, friends, consult, publications)
Obligation (Facebook, users, compose, identifiant)	Permission (u_3, friendOffriend, contact, account)
Permission (Facebook, users, comment, wall)	Permission (u_1, friends, publish, wall)
Permission (Facebook, users, create, pages)	Prohibition (u_1, public, consult, publications)
Prohibition (Facebook, friend3, publishinmywall, comment)	Prohibition (u_1, friend1, consult, photos)
Permission (Facebook, users, accept, friend_requests)	Prohibition (u_2, public, consult, account)

For the Facebook-User policy, every user can create an account by entering his/her identity and login information such as full name, gender, age, etc. The user is then able to share messages with friends, publish photos and videos, join groups, create pages and events, etc. Users' posts and publications can be managed by account owners, consulted or commented on by other Facebook users. For a User-User policy, users are able to manage their posts and information, and can allow or deny their friends, family or public users to access their data. By using the MotOrBAC simulator, the organization (Facebook) is defined, Facebook abstract and concrete entities, subjects and their association to roles, the context and permissions at an abstract level are also defined. Figure 4 illustrates the roles and definition of Facebook. Figure 5 indicates the generation of permissions at a concrete level, which are automatically generated by MotOrBAC (using *update* tool).

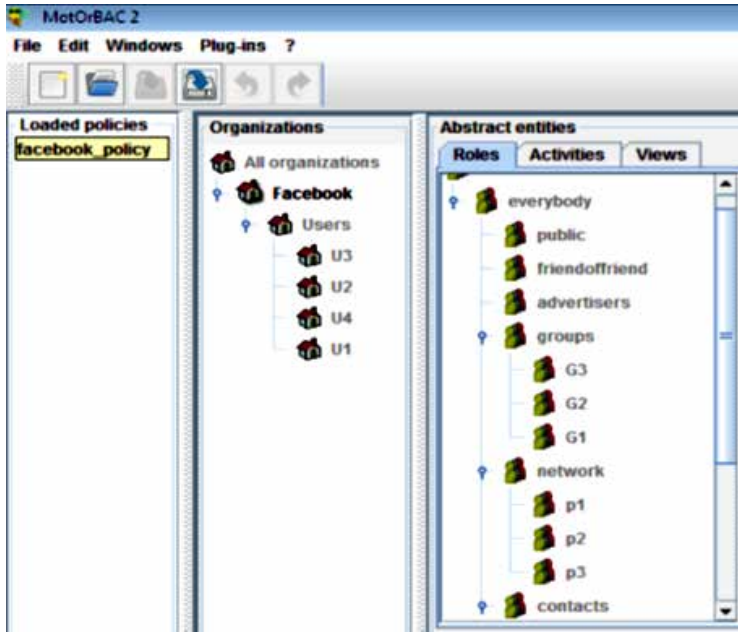


Figure 4: The roles and definition of Facebook using MotOrBAC

(Belbergui et al. 2016)

Derives from	Subject	Action	Object
permission6	Aimee	search	statut1
permission13	David	change	email_address
permission6	Aimee	see	site
permission2	Alexander	search	pictures
permission6	Aimee	view	number
permission13	David	change	identif_video1
permission6	Aimee	search	March

Figure 5 : The generation of permissions at the concrete level

(Belbergui et al. 2016)

The detection of policy coherence then follows and counts the conflicts in abstract and concrete levels, here are some examples :

- Example 1 : Conflict between permissions and prohibitions defined by two users :

  Prohibition (u_1, everyone, consult, relation u_1_u_4)

  Permission (u_4, everyone, consult, relation u_1_u_4)

Users u_1 and u_4 are friends, when u_1 prohibits everyone to consult this friendship and u_4 allows it, this generates a conflict. Such conflicts show that Facebook does not suggest any solution to these types of conflicts for users.

- Example 2 : Conflict between Facebook permissions and user u_1 prohibitions :

  Prohibition (u_1, advertisers, publishinmywall, publications)

  Permission (Face, advertisers, publishinmywall, publications)

Although user u_1 chooses not to publish advertisements on his wall, Facebook obliges him to be contacted by advertisers.

- Example 3 : Conflict between permissions and prohibitions assigned by Facebook to users.

  Prohibition (Face, P3, AccessControl, profile_photo)

  Permission (Face, P3, AccessControl, photos)

In this example, Facebook authorizes users to control the access to all of their photos except to their profile photo, which is always public.

Hence, OrBAC extension, which is adapted to Facebook, is used to analyze the coherence/incoherence of the Facebook security policy to enhance privacy features.

4.2.2. Other Access Control Methods

Multiparty Access Control Model

Social media networks provide virtual zones or spaces for users which are identified by their profile information and contain a list of friends for each user, web pages or walls. Although these networks provide some AC mechanisms that allow users to manage access to their own information within their individual space, they do not provide control over information that exists outside their own space. For this purpose, Hu et al. (2012b) propose a Multiparty Access Control (MPAC) Model to address the following issues:

- Users can post comments on their friends' spaces (or wall), but they cannot specify which users can view them. They can tag friends by uploading photos, but the tagged friends are unable to control who can see these photos and privacy concerns may be an issue.
- Social networks provide primitive protection mechanisms for these issues, such as:
  - allowing tagged users (e.g. Facebook) to remove the tags linked to their profiles;
  - allowing users to report violations to social network site managers by requesting for removal of the content they refuse to share with the public.

These mechanisms have several limitations as mentioned by Hu et al., for example, a tagged user's image is still discovered by all users who are authorized by the user who tags it even when a tag is removed from a photo. For this reason, the MPAC model is proposed as a solution to facilitate collaborative management of shared data in social networks in (Hu et al. 2012b). In MPAC, three scenarios are analyzed: profile sharing, relationship sharing, and content sharing.

- In profile sharing, social applications consume user profile attributes, such as name, birthday, activities, etc. of a user's friends. Figure 6 depicts MPAC Pattern for profile sharing where users are allowed to select some of their profile attributes to share with the applications when their friends use these applications. The user's friend is the owner of shared profile attributes, the application is

an accessor, and the user is a disseminator. Hence, a disseminator is able to share others' profile attributes to an accessor, and together, the owner and the disseminator can specify AC policies to restrict the sharing of profile attributes.

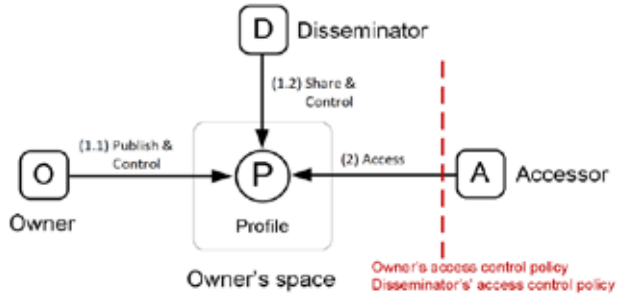


Figure 6 : MPAC pattern for profile sharing

(Hu et al. 2012)

- Relationship sharing is where users can share their relationships with other members and these relationships are inherently bidirectional and might carry some sensitive information. Most social networks allow users to control their friends list display, in this case a user can only manage one direction of a relationship. Figure 7 illustrates the scenario of a relationship sharing pattern. The owner which refers to the user and has a relationship with another user called stakeholder, shares the relationship with an accessor. Hence, authorization requirements should be considered from the stakeholder and the owner, as privacy concerns for the stakeholder may be violated.

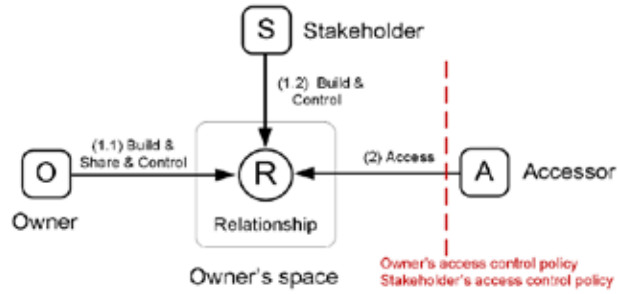


Figure 7 : MPAC pattern for relationship sharing

(Hu et al. 2012)

- In content sharing, users can communicate and share contents with other members. Social network users can tag or share others to the contents they upload or post on their pages. These contents may be related or connected with multiple users. In this scenario, three examples are explained to represent MAPC Pattern for content sharing and are shown in Figure 8. The first example is illustrated in Figure 8 (a), user A uploads a photo which also contains other friends B and C. User A is the owner of the photo, B and C are stakeholders of it. In this case, all users can determine AC policies to control who can see this photo, not only the owner. In other words, the content has many stakeholders who can be involved in the control of content sharing. The second is illustrated in Figure 8 (b), when user A posts a message on B's wall mentioning user C. In this case, user A is called a contributor of the message, user C is identified by a mention and considered as a stakeholder of the message, hence users A and C may want to control the disclosure of this message. As shown in Figure 8 (b), a contributor (user A) publishes a content to others' wall and this content might have many stakeholders or tagged users. In this case, all related users should be authorized to define AC policies for the posted content. The third example is demonstrated in Figure 8 (c), where users are allowed to share others' content in such a way where user A shares content (e.g. a photo) with his friends after viewing it in user B's wall. The shared photo is now in user A's space and he can determine AC policy to allow/deny his friends to see this photo. In this situation, user A is a photo disseminator and his privacy concerns might differ from that of user B. This could cause leakage of sensitive information via the data dissemination procedure. Hence, in Figure 8 (c), the owner or the contributor shares his content by uploading and publishing it, then the disseminator is able to view and share this content. In this case, to regulate content access in the disseminator's space, all AC policies that are defined by associated users must be enforced.

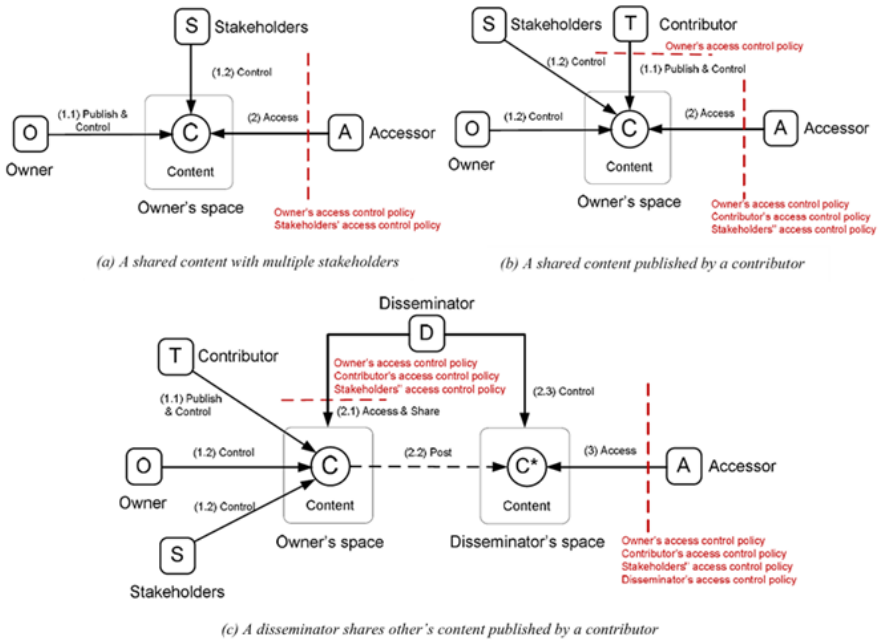


Figure 8 : MPAC pattern for content sharing

(Hu et al. 2012)

In a MPAC system, a group of users can collaborate together to influence the final AC decision.

PACMAN : Personal Agent for Access Control in Social Media

Personal Agent for AC in Social Media (PACMAN) is proposed by Misra et al. (2017) as a personal assistant agent that recommends personalized AC decisions on any information disclosure on social environments. This can be done by combining groups generated from the user's network structure and using information in the user's profile. Since social media users do plenty of interactions, an appropriate mechanism is needed to control information access by selecting the appropriate audience or friends from their lists. PACMAN is presented as a personal agent for a user to calculate accurate recommendations and minimize obtrusiveness. Figure 9 illustrates PACMAN components and inputs to produce an AC recommendation ("allow" or "deny").

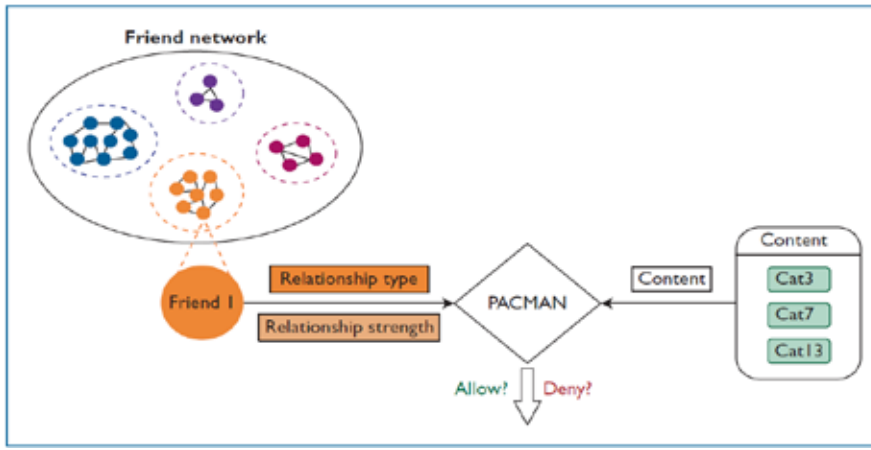


Figure 9 : Components and inputs of PACMAN

(Misra et al., 2017)

Relationship types are the interpersonal interactions between friends, colleagues, family, etc. and social media users. Relationship strength is the strength or closeness of interpersonal relationships between social media users. This strength is estimated by measuring similarities between users' profiles. For this purpose, various methods are proposed for estimating the relationships tie-strength or closeness (Fogués et al. 2014, Misra et al. 2016a). Users' total friends and mutual friends, as stated by Misra et al. (2017), are the most suitable profile attributes to support prediction of AC decisions. Moreover, the content of information that is being shared is used to enhance AC methods. To address this issue, various methods are used to generate attributes depending on the nature of the content, "for example, natural language processing techniques can be used for text, and image processing can be used for photos (Misra et al. 2017)".

For PACMAN implementation, several building blocks are used (Figure 9), and the user's friend network is required as an input. To represent the relationship type, one of the network-based community detection algorithms used is called Clique Percolation Method (CPM) (Misra et al. 2016b). For relationship strength, total friends and mutual friends, which are fetched from the users' profile, are also used as input to the PACMAN mechanism. For the type of content being shared, various methods to obtain content information can be applied, Misra et al. (2017) use "manual selection of photo categories in the form of "tags" to represent the information about content". Consequently, "allow" or "deny" AC decisions

to the user which are recommended by PACMAN are of equal importance, this reflects the importance of accuracy in this mechanism. Accuracy is calculated as a percentage of the total recommendations that are correct, where :

$$\text{Accuracy} = ((F - \text{Errors})/F).$$

F is the number of total friends for a user. Errors include allow and deny errors, such errors as stated in (Misra et al. 2017) arise in the following cases :

- "An *allow* error occurs when PACMAN recommends a *deny* decision to the user when it actually should have been *allowed*. These errors are essentially *false negative (FN)* recommendations and result in a *deny* to *allow* change being made by the user."
- "A *deny* error occurs when PACMAN recommends an *allow* decision to the user when it actually should have been *denied*. These errors are *false positive (FP)* recommendations and result in an *allow* to *deny* change by the user."

Hence, Errors = FN + FP

For the experiment, an application similar to Facebook is created using Facebook Query Language (FQL), and a sample of 26 participants are asked to upload 10 photos (per user). The users are then asked to select categories for the photos in the form of tags to represent the content information. To calculate accuracy of prediction produced for each individual user, Weka is integrated into PACMAN to create and run the classifier applying 10-fold cross validation. Thereafter, accuracy of recommendations produced by PACMAN are shown in Figure 10.

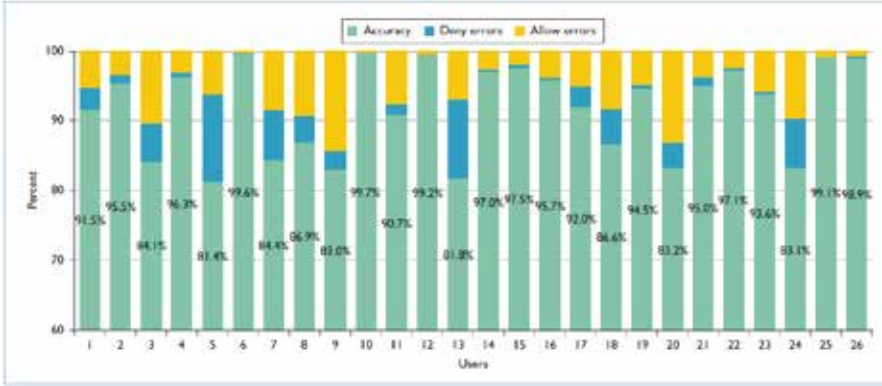


Figure 10 : Accuracy and ratio of changes required to recommendations made by PACMAN for all 26 users

(Misra et al. 2017)

The ratio of incorrect recommendations, *allow* and *deny* errors, shows that PACMAN produces good quality recommendations since highly accurate recommendations are demonstrated for almost all users.

5. CONCLUSION

Access control methods are used in computing environments to mitigate security and privacy risks of unauthorized and illegal access to data. These methods vary depending on the underlying structure of the system environment and the needed level of protection. In this chapter, a spacious overview of the main aspects of AC methods as solutions to various privacy and security related issues in social media networks is provided. First, the social media network types and services, the possible threats, and the main privacy problems in these networks are reviewed. Then, the importance of AC methods and the essential requirements for social networks are explained. Subsequently, the common AC methods that are used as a basis and implemented in different computing environments are summarized. Consequently, we present the state-of-the-art for some recent AC methods for social networks, and in the description of each presented method, we highlight the main contribution of the model with the different approaches.

Based on the aforementioned work we find that, although social media networks have a set of privacy policies, they are vulnerable to various kinds of attacks and privacy issues. The kind of personal data in these networks

needs a high level of privacy protection by means of appropriate access control. Some AC methods are proposed in this domain to tackle the particular structure and the fundamental privacy issues of social networks, and some other AC methods are proposed and dedicated only for some social network sites such as Facebook (Anwar et al. 2010) and Google+ (Hu et al. 2012a). In this context, is it possible to find an AC method that works as a general basis and include all the needed features to enforce AC policy in social network sites, since all the presented methods and despite their different mechanisms focus on the same privacy issues for social media users. Furthermore, the proposed AC methods reflect that finding AC methods for social network users is a recent research issue, and research is still being conducted due to the lack of privacy features of social network sites, especially that social networks are dynamically changing environments.

REFERENCES

Aldhafferi, Nahier, Charles Watson, and AS Sajeev. 2013. "Personal information privacy settings of online social networks and their suitability for mobile internet devices." *arXiv preprint arXiv :1305.2770*.

Ali, Shaukat, Naveed Islam, Azhar Rauf, Ikram Din, Mohsen Guizani, and Joel Rodrigues. 2018. "Privacy and Security Issues in Online Social Networks." *Future Internet* 10 (12) :114.

Anwar, Mohd, Zhen Zhao, and Philip WL Fong. 2010. An access control model for Facebook-style social network systems. University of Calgary.

Ausanka-Crues, Ryan. 2001. "Methods for access control : advances and limitations." *Harvey Mudd College* 301 :20.

Belbergui, Chaimaa, Najib Elkamoun, and Rachid Hilal. 2016. "Modeling Access Control Policy of a Social Network." *International Journal of Advanced Computer Science and Applications* 7 (6).

Belokosztolszki, András. 2004. Role-based access control policy administration. University of Cambridge, Computer Laboratory.

Bin Jeffry, Mohd Aliff Faiz, and Hazinah Kutty Mammi. 2017. "A study on image security in social media using digital watermarking with metadata." 2017 IEEE Conference on Application, Information and Network Security (AINS).

Carminati, Barbara, Elena Ferrari, and Andrea Perego. 2006. "Rule-based access control for social networks." OTM Confederated International Conferences" On the Move to Meaningful Internet Systems".

Crampton, Jason. 2003. "On permissions, inheritance and role hierarchies." Proceedings of the 10th ACM conference on Computer and communications security.

Cutillo, Leucio Antonio, Mark Manulis, and Thorsten Strufe. 2010. "Security and privacy in online social networks." In *Handbook of Social Network Technologies and Applications*, 497-522. Springer.

Delerue, Helene, and Wu He. 2012. "A review of social media security risks and mitigation techniques." *Journal of Systems and Information Technology*.

Deliri, Sepideh, and Massimiliano Albanese. 2015. "Security and privacy issues in social networks." In *Data Management in Pervasive Systems*, 195-209. Springer.

Ennahbaoui, Mohammed, and Said Elhajji. 2013. "Study of access control models." Proceedings of the World Congress on Engineering.

Fiesler, Casey, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, and Munmun De Choudhury. 2017. "What (or who) is public? : Privacy settings and social media content sharing." Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.

Fire, Michael, Roy Goldschmidt, and Yuval Elovici. 2014. "Online social networks : threats and solutions." *IEEE Communications Surveys & Tutorials* 16 (4) :2019-2036.

Fogués, Ricard L, Jose M Such, Agustin Espinosa, and Ana Garcia-Fornes. 2014. "BFF : A tool for eliciting tie strength and user communities in social networking services." *Information Systems Frontiers* 16 (2) :225-237.

Ghazinour, Kambiz, Stan Matwin, and Marina Sokolova. 2016. "YOURPRIVACYPROTECTOR, A recommender system for privacy settings in social networks." *arXiv preprint arXiv :1602.01937*.

Gupta, Brij B, Nalin AG Arachchilage, and Kostas E Psannis. 2018. "Defending against phishing attacks : taxonomy of methods, current issues and future directions." *Telecommunication Systems* 67 (2) :247-267.

Hu, Hongxin, Gail-Joon Ahn, and Jan Jorgensen. 2012a. "Enabling collaborative data sharing in google+." 2012 IEEE Global Communications Conference (GLOBECOM).

Hu, Hongxin, Gail-Joon Ahn, and Jan Jorgensen. 2012b. "Multiparty access control for online social networks : model and mechanisms." *IEEE Transactions on Knowledge and Data Engineering* 25 (7) :1614-1627.

Hu, Vincent C., David F. Ferraiolo, Ramaswamy Chandramouli, and D. Richard Kuhn. 2017. *Attribute-Based Access Control* Norwood : Artech House.

Ikhalia, E, and CO Imafidon. 2013. "The need for two factor authentication in social media." Proceedings of the International Conference on Future Trends in Computing and Communication-FTCC.

Jahid, Sonia, Prateek Mittal, and Nikita Borisov. 2011. "EASiER : Encryption-based access control in social networks with efficient revocation." Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security.

Jain, Sakshi, Juan Lang, Neil Zhenqiang Gong, Dawn Song, Sreya Basuroy, and Prateek Mittal. 2015. "New directions in social authentication." Proc. USEC.

Jin, Xin, Ram Krishnan, and Ravi Sandhu. 2012. "A unified attribute-based access control model covering DAC, MAC and RBAC." IFIP Annual Conference on Data and Applications Security and Privacy.

Joe, M Milton, and B Ramakrishnan. 2017. "Novel authentication procedures for preventing unauthorized access in social networks." *Peer-to-Peer Networking and Applications* 10 (4) :833-843.

Kashmar, Nadine, Mehdi Adda, and Mirna Atieh. 2019. "From Access Control Models to Access Control Metamodels : A Survey." Future of Information and Communication Conference.

Kashmar, Nadine, Mehdi Adda, Mirna Atieh, and Hussein Ibrahim. 2019a. "A new dynamic smart-AC model methodology to enforce access control policy in IoT layers." Proceedings

of the 1st International Workshop on Software Engineering Research & Practices for the Internet of Things.

Kashmar, Nadine, Mehdi Adda, Mirna Atieh, and Hussein Ibrahim. 2019b. "Smart-AC : A New Framework Concept for Modeling Access Control Policy." *Procedia Computer Science* 155 :417-424.

Kashmar, Nadine, Mehdi Adda, Mirna Atieh, and Hussein Ibrahim. 2020. "Deriving Access Control Models based on Generic and Dynamic Metamodel Architecture : Industrial Use Case." The 11th International Conference on Emerging Ubiquitous Systems and Pervasive Networks. *(Accepted)*

Kayem, Anne VDM, Selim G Akl, and Patrick Martin. 2010. *Adaptive cryptographic access control*. Vol. 48 : Springer Science & Business Media.

Lee, Sangho, and Jong Kim. 2013. "Warningbird : A near real-time detection system for suspicious urls in twitter stream." *IEEE transactions on dependable and secure computing* 10 (3) :183-195.

Li, Fengyong, Kui Wu, Jingsheng Lei, Mi Wen, Zhongqin Bi, and Chunhua Gu. 2015. "Steganalysis over large-scale social networks with high-order joint features and clustering ensembles." *IEEE Transactions on Information Forensics and Security* 11 (2) :344-357.

Madejski, Michelle, Maritza Johnson, and Steven M Bellovin. 2012. "A study of privacy settings errors in an online social network." 2012 IEEE International Conference on Pervasive Computing and Communications Workshops.

Miller, Zachary, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. 2014. "Twitter spammer detection using data stream clustering." *Information Sciences* 260 :64-73.

Misra, Gaurav, and Jose M Such. 2017. "Pacman : Personal agent for access control in social media." *IEEE Internet Computing* 21 (6) :18-26.

Misra, Gaurav, Jose M Such, and Hamed Balogun. 2016a. "IMPROVE-Identifying Minimal PROfile VEctors for similarity based access control." 2016 IEEE Trustcom/BigDataSE/ISPA.

Misra, Gaurav, Jose M Such, and Hamed Balogun. 2016b. "Non-sharing communities ? an empirical study of community detection for access control decisions." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

O'Malley, A James, and Jukka-Pekka Onnela. 2017. "Introduction to social network analysis." *Methods in Health Services Research* :1-44.

Patsakis, Constantinos, Athanasios Zigomitros, Achilleas Papageorgiou, and Agusti Solanas. 2015. "Privacy and security for multimedia content shared on OSNs : issues and countermeasures." *The Computer Journal* 58 (4) :518-535.

Rathore, Shailendra, Pradip Kumar Sharma, Vincenzo Loia, Young-Sik Jeong, and Jong Hyuk Park. 2017. "Social network security : Issues, challenges, threats, and solutions." *Information sciences* 421 :43-69.

Sachan, Amit, and Sabu Emmanuel. 2011. "Efficient Access Control in Multimedia Social Networks." In *Social Media Modeling and Computing*, 145-165. Springer.

Sandhu, Ravi, David Ferraiolo, and Richard Kuhn. 2000. "The NIST model for role-based access control : towards a unified standard." ACM workshop on Role-based access control.

Sayaf, Rula, and Dave Clarke. 2014. "Access control models for online social networks." In *Digital Arts and Entertainment : Concepts, Methodologies, Tools, and Applications*, 451-484. IGI Global.

Taleby Ahvanooey, Milad, Qianmu Li, Jun Hou, Ahmed Raza Rajput, and Chen Yini. 2019. "Modern text hiding, text steganalysis, and applications : a comparative analysis." *Entropy* 21 (4) :355.

Tapiador, Antonio, Diego Carrera, and Joaquín Salvachúa. 2012. "Tie-RBAC : an application of RBAC to Social Networks." *arXiv preprint arXiv :1205.5720*.

Zhang, Chi, Jinyuan Sun, Xiaoyan Zhu, and Yuguang Fang. 2010. "Privacy and security for online social networks : challenges and opportunities." *IEEE network* 24 (4) :13-18.

Zhang, Zhiyong, and Brij B Gupta. 2018. "Social media security and trustworthiness : overview and new direction." *Future Generation Computer Systems* 86 :914-925.

Zigomitros, Athanasios, Achilleas Papageorgiou, and Constantinos Patsakis. 2012. "Social network content management through watermarking." 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications.

5

SOCIAL MEDIA SURVEILLANCE : BETWEEN DIGITAL GOVERNMENTALITY, BIG DATA AND COMPUTATIONAL SOCIAL SCIENCE

Ramón Reichert

Ramón Reichert is a research assistant professor at the Department of Art and Education at the University of Art and Design in Linz. He currently works as a Senior Researcher and EU-project coordinator for the research project Addressing Violent Radicalisation : A Multi-actor Response through Education, that is ISF-P funded. He is the program director of the M.Sc. Data Studies at Danube University Krems, Austria. He is a lecturer at the Department of Art Sciences and Art Education at the University of Applied Arts Vienna, and a lecturer in Contextual Studies at the School of Humanities and Social Sciences at the University of St. Gallen, Switzerland. He works as a researcher with a particular focus on media change and social changes in the fields of theory and history of digital media, history of knowledge and media history of digital cultures, and media aesthetics. He is the co-editor of the refereed, international journal Digital Culture & Society and he is author of New Media Reader (2007, co-edited), Amateurs in the Internet Age (2008), Big Data (2014, ed.), Digital Material/ism (2015, co-edited), Rethinking AI : Neural Networks, Biometrics and the New Artificial Intelligence (2018, co-edited), Digital Citizens (2019, co-edited), Social Machine Facebook (2019, co-edited) and Selfie Culture (2019).

In his influential book *The Order of Things : An Archeology of the Human Sciences* (1966), Michel Foucault uses the term "episteme" to refer to the archeology of knowledge that is asserted in different periods of science history. In his analysis of historical knowledge and power-knowledge Foucault references three epistemes. The episteme of the Renaissance is characterized by similarity and affinity. Representation and their main structures, categorization, mathesis and taxonomy, is the main part of the classical episteme. Within the modern episteme, the relationship to reality is detached and the things cannot longer be understood as clearly identifiable and categorizable entities. What marks the digital episteme against this background ? What type of knowledge is created today in the digital communication spaces ? Which technical factors are responsible for designing knowledge in digital media environments ? The continuation of this theoretical framework raises the following epistemological question : What knowable and perceptible made digital networking media possible and which status do they have in the formation of society and culture ?

Google, Microsoft, Apple, Facebook - just about every company that designs software and builds digital infrastructures, such as data centers and server farms, hopes that the processing of ever-larger and more differentiated data sets will improve their understanding of social reality world. Keywords such as Big Data and New Artificial Intelligence not only rewrite new scientific data practices, but also mark profound transformations of contemporary society and a media culture in the digital upheaval. A number of theoretical reflections on digital societies assume that online platforms and social media are becoming a dominant media channel for participatory engagement. (Dijck 2012 ; Helmond 2015) Practices of participation and engagement are an important part of our digital everyday lives : from chat rooms to community forums, from social media platforms to image boards, and from rating platforms to whistle-blowing websites. (Papacharissi 2015)

Popular online services such as Facebook, Twitter, Whatsapp, Instagram or Youtube are used to produce participatory environments to build protest cultures and civil society engagement. (Fuchs 2014 : 52-63) It is often forgotten that social media are market-oriented providers. (Trottier/Fuchs 2015, Lyon 2017) By not reflecting the medium as a productive force itself, the asymmetric infrastructures between media usage and usage analysis are also overlooked. (Degli Esposti 2014 : 209-225) Capitalist social media are characterized by a contradiction : they are companies that monitor and exploit their users, while at the same time trying to keep their financial

flows and corporate structures secret in order to increase their profits through tax avoidance and monopolistic structures. (McChesney 2013 ; Scholz 2016)

Although the network structures of the peer-to-peer networks enable opportunities for action and cooperation beyond the legitimized institutions, they also remain vulnerable to micropolitics and personal exercise of power (Helmond 2015). The general dissolution of power and domination manifests itself in the organization of networks involved in processes of economization and internal marketization. The dominant class of digital hyper-workers is made up of the virtuosos of networking and the self-marketing professionals, who place those socially marginalized who are unable to assert themselves on the attention markets by means of popular cultural distinction work, fashion and lifestyles.

Studies on the political economy of social media have shown that manufacturing platform-content is linked with a market-oriented economy that is monopoly-managed. In my presentation, I would therefore like to show to what extent production relations between the advertising-financed platform owners and the platform users are asymmetrical, insofar as the users increase the value of the platform through their activities, they can't design or change the infrastructural order of the platform itself. These limitations of the means of production and the division of digital means of production mark an asymmetry between digital ownership and distributed work. (Scholz 2012) In this context, I would like to emphasize that discursive publics on online platforms are a digital construct of the algorithmic control by also subjecting political and social discourses to the logic of market-like trend settings and preference models of consumer research as described by Jordan (2015), Means (2018) and Fuchs (2018 : 677-702).

Learning algorithms and predictive models are part of social media applications and create new epistemic conditions for digital biometrics. Using (complete) autonomous agents, New Artificial Intelligence examines certain issues and basic concepts such as "self-sufficiency", "autonomy and situatedness", "embodiment" "adaptivity" and "ecological niches and universality" involving vast areas of human and social sciences. In this context, Kate Crawford (Microsoft Research) has recently warned against the impact that current AI research might have, in a noteworthy lecture titled : AI and the Rise of Fascism. Crawford analyzed the risks and potential of AI research and asked for a critical approach in regard to new forms of

data-driven governmentality : "Just as we are reaching a crucial inflection point in the deployment of AI into everyday life, we are seeing the rise of white nationalism and right-wing authoritarianism in Europe, the US and beyond. How do we protect our communities – and particularly already vulnerable and marginalized groups – from the potential uses of these systems for surveillance, harassment, detainment or deportation ?" (Crawford 2017)

Against this background, I would like to examine the structural connection between digital monopoly capitalism and the asymmetry between usage architectures (frontend) and data exploitation (backend). In doing so, I deal with the relevant data and power-critical positions of Surveillance Studies and Critical Software Studies in order to develop problem-oriented further developments of these theory models.

In the era of big data, not only the importance of social but also of scientific knowledge has radically changed. Social media, mobile devices and technical assistance systems today function as gigantic data collectors and as relevant data sources of digital communication research : "Social media offers us the opportunity for the first time to observe human behavior and interaction in real time and on a global scale." Golder / Macy 2012 : 7) The installation of media surveillance systems is not a new phenomenon. Panopticon surveillance architectures to public video surveillance (CCTV) play with the moment of internalizing surveillance ("how can I be sure I'm not being monitored, so I behave predictably to avoid attracting attention"), but digital networking media are essentially based on active self-monitoring, which must be made evident in order to be recognized.

This principle of control not only communicates with the supervisor, but it also seeks to maximize the principle of supervision. Surveillance is most effective when as many as possible can monitor at the same time. In this sense, social networking media can be seen as a place of mutual and ongoing control. While visibility at Foucault was still a trap for the power of an observer, the social media panopticon is all about the continued multiplication of visibility and visual control - that's the big difference to CCTV, which deletes observation to experts.

The large amounts of data are collected in a variety of fields of knowledge : biotechnology, genome research, labor and finance sciences and trend research rely on the results of big data, data processing and formulate meaningful models on the current status and future development

of social groups and societies. Digital mass data research has become considerably differentiated in recent years, and numerous studies have been published that use computer-based methods such as text analysis (quantitative linguistics), sentiment analysis (sentiment recognition), social network analysis or image analysis of other machine-based methods of computer-based social media -Analysis.

In the era of big data, the importance of social networking culture has changed radically. Social media acts today as a gigantic data collector and as a relevant data source fore digital communications research : "Social media offers us the opportunity for the first time to both observe human behavior and interaction in real time and on a global scale." (Golder/Macy 2012 : 7) The large amounts of data that are collected in different domains of knowledge, such as biotechnology, genomics, social science, health care and financial sciences or trend research, rely, in their work and studies, on the results of information processing of big data and formulate on this basis, significant models of the current status and future development of social groups and societies. Big data research became significantly differentiated in recent years, as numerous studies have been published using machine-based methods such as text analysis (quantitative linguistics), sentiment analysis (mood detection), social network analysis, and image analysis or machine-based processes of computer-based social media analysis.

The newly emerging discipline of Computational Social Science (Lazer et al, 2009 : 721-723 ; Conte et al.2012 : 325-346) evaluates the large amounts of online data use in the backend area and has emerged as a new leading science in the study of social media web 2.0. It provides a common platform for computer science and social sciences connecting the different expert opinions on computer science, society and cultural processes. Computer science deals with computer-based elaboration of large databases, which no longer cope with the conventional methods of statistical social sciences. Its goal is to define the social behavioral patterns of online users based on methods and algorithms of data mining : "To date, research on human interactions has relied mainly on one-time, self-reported data on specific relationships between usage and behavior." (Lazer et.al. 2009 : 722) In order to answer the question of social behavior in a relevant or meaningful way, computer science requires the methodological input of social sciences. With their knowledge of social activity theories and methods , social sciences make a valuable contribution to the formulation of relevant issues.

In the last few years, however, Critical Code Studies have resulted in a research network that would allow digital meshing technologies a structure-forming power and form a productive interface for transdisciplinary data criticism. Assuming that collective data communication in computer networks is organized through the network infrastructure of network protocols and dissected into functional Internet layers, the network protocols can be viewed as cultural techniques of social regulation, using collective processes as technological effects of network technologies (Wagner 2006 : 26-27).

Influential theorists have warned of a "digital divide", in which knowledge about the future is unequally distributed, potentially leading to imbalances of power between researchers inside and outside networks. Lev Manovich argues that the limitation of access to social data creates a monopoly in the government and administration of the future : "Only social media companies have access to really large social data – especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not." (Manovich 2012 : 467)

This inequality cements the position of the social networks as computer-based media of control, whose knowledge is acquired via a vertical and one-dimensional web of communication. The social networks enable a continual flow of data (digital footprints), which they collect and organize, establishing closed spaces of knowledge and communication for experts, who distil the data into information. Knowledge about the future thus passes through various technical and infrastructural levels that are organized hierarchically and in pyramidal order. Clearly, alongside the technical-infrastructural isolation of knowledge about the future, strategic decision-making is also located at the back end rather than in peer-to-peer communication. While peers are able to falsify results, create fake profiles and communicate nonsense, their limited agency prevents them from moving beyond these tactics and actively shaping the future.

In their approach to software architecture and social networks, Alexander Galloway and Eugene Thacker (2007) deepen the question of the emergence of systemic data as a social institution, highlighting the importance of computer science concepts and user interfaces in creating social formations and political figures of knowledge. They interpret the algorithmic standards, norms and protocols as a mediating authority between cultural practices and technical infrastructure. They understand

networks not just as technical systems, but as socially dynamic and vibrant networks that organize themselves in real time. Against this background, they not only examine the technical possibilities of political control through algorithms and protocols, but also question the political options for action of network-based movements.

ALGORITHMIC GOVERNMENTALITY

At the interface between the "computational social science" (Lazer et al.2009) and the "cultural analytics" (Manovich, 2009 : 199-212) an interdisciplinary theoretical field has emerged, reflecting the new challenges of the digital Internet research. The approach of computational surveillance aims to rethink the research about use (audience research), by interpreting the use practices of the Internet as a cultural change and as social issues. (Rogers 2011 : 63) Analogous methods though, that have been developed for the study of interpersonal or mass communication, cannot simply be transferred to digital communication practices. Digital methods can be understood as approaches that focus on the genuine practice of digital media, and not on existing methods adapted for Internet research. According to Rogers (2011), digital methods are research approaches that take advantage of large-scale digital communication data to, subsequently, model and manage this data using computational processes. Both the approach of "computational social science" and the questioning of "digital methods" represent the fundamental assumption that by using the supplied data, which creates social media platforms, new insights into human behavior, into social issues beyond these platforms and into their software can be achieved. Numerous representatives of computer-based social and cultural sciences sustain the assumption that online data could be interpreted as social *environments*. To do so, they define the practices of Internet use by docking them using a positivist data term, which comprehends the user practices as an expression of specifiable social activity. The social positivism of "computational social science" in social media platforms neglects, however, the meaningful and intervening/ instructive role of the media in the production of social roles and stereotyped conducts in dealing with the medium itself. With respect to its postulate of objectivity, social behaviorism of online research can, in this regard, be questioned.

The vision of such native-digital research methodology, whether in the form of a "computational social science" (Lazer 2009 : 721-723) or "cultural

analytics" (Manovich, 2009 : 199-212) is, however, still incomplete and requires an epistemic survey of digital methods in algorithmic governmentality of the following areas :

(1) Algorithmic governmentality as *validity theoretical project* stands for a specific process that claims the social recognition of action orientations. The economy of computer science, computational linguistics and empirical communication sociology not only form a network of scientific fields and disciplines, but they also develop, in their strategic collaborative projects, certain expectations, describing and explaining the social world and are, in this respect, intrinsically connected with epistemic and political issues. In this context, the epistemology, questioning the self-understanding of digital methods, deals with the social effectiveness of the digital data science.

(2) Algorithmic governmentality as *constitutional theoretical construct*. The relation to the object in big data research is heterogeneous and consists of different methods. Using interface technologies, the process of data tracking, of keyword tracking, of automatic network analysis, of argument and sentiment analysis or machine-based learning, results in critical perspectivizations of data constructs. Against this background, the *Critical Code Studies* try to make the media techniques, of computer science power relations, visible and study the technical and infrastructural controls over layer models, network protocols, access points and algorithms.

(3) Algorithmic governmentality may ultimately be regarded as a *founding theoretical fiction*. The relevant research literature has dealt extensively with the reliability and validity of scientific data collection and came to the conclusion that the data interfaces of Social Net (Twitter, Facebook, YouTube) act more or less like dispositive orders according to a gatekeeper. The filter interface generates the APIs (application programming interfaces), economically motivated exclusionary effects for network research that cannot be controlled by their own efforts. In this context of problem-oriented development of computer science, the expectations on the science of the 21st century have significantly changed. In the debates increasingly claims are being made, they insist in processing historically, socially and ethically leading aspects of

digital data practices associated with the purpose to anchor these aspects in the future scientific cultures and epistemologies of data generation and data analysis. Lazer et. al. demand of future computer scientists a responsible use of available data and see in negligent handling a serious threat to the future of the discipline itself : "A single dramatic incident involving a breach of privacy could produce a set of statutes, rules, and prohibitions that could strangle the nascent field of computational social science in its crib. What is necessary, now, is to produce a self-regulatory regime of procedures, technologies, and rules that reduce this risk but preserve most of the research potential." (Lazer et. al. 2009 : 722) If it is going to be made research on social interaction using computer science and big data, then the responsible handling of data as well as the compliance with data protection regulations are key issues. In the last part of my essay, I want to look at the relationship between algorithmic-technical control, participatory government and biometric measurement of users of online platforms and social networking sites.

DIGITAL BIOSURVEILLANCE

The digital measurement of biometric activity is one of the most popular and widespread power practices of monitoring digital usage cultures. Before this background the so-called internet biosurveillance, or in other words the digital disease detection, represents a new paradigm of general issue of the public health governance. While traditional approaches to health prognosis operated with data collected in the clinical diagnosis, the Internet biosurveillance studies use the methods and infrastructures of health informatics. This means, more precisely, that they use unstructured data from different web-based sources and targets using the collected and processed data and information about changes in health-related behavior. The two main tasks of the internet biosurveillance are (1) the early detection of epidemic diseases, biochemical, radiological and nuclear threats (Brownstein 2009) and (2) the implementation of strategies and measures of sustainable governance in the target areas of health promotion and health education. (Walters, 2010) Biosurveillance has established itself as an independent discipline in the mid-1990s, as military and civilian agencies began to get interested in automatic monitoring systems. In this context,

the biosurveillance program of the Applied Physics Laboratory of Johns Hopkins University has played a decisive and pioneering role. (Burkom, 2008)

The internet biosurveillance uses the accessibility to data and analytic tools provided by digital infrastructures of social media, participatory sources and non-text-based sources. The structural change generated by digital technologies, as main driver for big data, offers a multitude of applications for sensor technology and biometrics as key technologies. Biometric analysis technologies and methods are finding their way into all areas of life, changing people´s daily lives. In particular, the areas of sensor technology, biometric recognition process and the general tendency towards convergence of information and communication technologies are stimulating the big data research. The conquest of mass markets through sensor and biometric recognition processes can sometimes be explained by the fact that mobile, web-based terminals are equipped with a large variety of different sensors. More and more users come this way into contact with the sensor technology or with the measurement of individual body characteristics. Due to the more stable and faster mobile networks, many people are permanently connected to the internet using their mobile devices, providing connectivity an extra boost.

With the development of apps, application software for mobile devices such as smartphones (iPhone, Android, BlackBerry, Windows Phone) and tablet computers, the application culture of biosurveillance changed significantly, since these apps are strongly influenced by the dynamics of the bottom-up participation. Andreas Albrechtslund speaks in this context of the "Participatory Surveillance" (2008) on the social networking sites, in which biosurveillance increasingly assumes itself as a place for open production of meaning and permanent negotiation, by providing comment functions, hypertext systems, ranking and voting procedures through collective framing processes. This is the case of the sports app Runtastic, monitoring different sports activities, using GPS, mobile devices and sensor technology, and making information, such as distance, time, speed and burned calories, accessible and visible for friends and acquaintances in real-time. The Eatery app is used for weight control and requires its users the ability to do self-optimization through self-tracking. Considering that health apps also aim to influence the attitudes of their users, they can additionally be understood as persuasive media of health governance. With their feedback technologies, the apps facilitate not only issues related to

healthy lifestyles, but also multiply the social control over compliance with the health regulations in peer-to-peer networks. Taking into consideration the network connection of information technology equipment, as well as the commercial availability of biometric tools (e.g. "Nike Fuel", "Fit Bit", "iWatch") and infrastructure (apps), the biosurveillance is frequently associated, in the public debates, to dystopian ideas of a society of control biometrically organized.

Organizations and networks for health promotion, health information, health education and formation observed with great interest that, every day, millions of user search for information about health using the Google search engine. During the influenza season the searches for flu increase considerably and the frequency of certain search terms can provide good indicators of flu activity. Back in 2006, Eysenbach evaluated in a study on "Infodemiology" or "Infoveillance" the Google *AdSense* click quotas, with which he analyzed the indicators of the spread of influenza and observed a positive correlation between increasing search engine entries and increased influenza activity. Further studies on the volume of search patterns have found that there is a significant correlation between the number of flu-related search queries and the number of people showing actual flu symptoms. (Freyer-Dugas, 2012) This epidemiological correlation structure was subsequently extended to provide early warning of epidemics in cities, regions and countries, in cooperation with the 2008 established *Google Flu Trends* in collaboration with the US authority for the surveillance of epidemics (CDC). On the *Google Flu Trends* website, users can visualize the development of influenza activity both geographically and chronologically. Some studies criticize that the predictions of the Google project are far above the actual flu cases.

Ginsberg et al. (2009) point out that in the case of an epidemic it is not clear whether the search engines behavior of the public remains constant, and thus whether the significance of *Google Flu Trends* is secured or not. They refer to the medialized presence of the epidemic as distorting cause of an "Epidemic of Fear" (Eysenbach, 2006 : 244), which can lead to miscalculations concerning the impending influenza activity. Subsequently, the prognostic reliability of the correlation between increasing search engine entries and increased influenza activity has been questioned. In recent publications on digital biosurveillance, communication processes in online networks are more intensely analyzed. Especially in the field of Twitter Research (Paul/Dredze, 2011), researchers developed specific

techniques and knowledge models for the study of future disease development and work backed up by context-oriented sentiment analysis and social network analysis to hold out the prospect of a socially and culturally differentiated biosurveillance.

SUMMARY AND OUTLOOK

In recent years social media and online platforms have become an important source database for mass statistical surveys, and they have generated new forms of social-empirical knowledge through data-driven digital methods. Its gigantic databases are used for gathering social information and are used for collecting, analyzing and interpreting social statistics and information. In their role as a storage, processing and dissemination medium of social mass data, social networks have produced extensive data aggregates that are used to predict societal developments. Their future knowledge overlays two fields of knowledge. Empirical social science and media informatics are responsible for the evaluation of mediated communication. Social research is evaluating the online communication media as a major force for social development. Against this background, it seems important to develop a substantial data-critical perspective, which has to develop a broader political theory of the data society that the thinking of power can be transferred into the digital present.

FURTHER READINGS

Albrechtslund, A. « Online Social Networking as Participatory Surveillance », in : *First Monday* vol. 13/3 (2008), Online : http ://firstmonday.org/ojs/index.php/fm/article/viewArticle/2142/1949

Albrechtslund, A. New media and changing perceptions of surveillance. In J. Hartley, J. Burgess & A. Bruns (eds.), *A companion to new media dynamics*, pp. 309-321. Oxford : Wiley-Blackwell 2013.

Burkom, H. S. et.al. « Decisions in Biosurveillance Tradeoffs Driving Policy and Research », in : *Johns Hopkins Technical Digest*, vol. 27/4 (2008), 299-311.

Brownstein, J. S. et.al. « Digital disease detection—harnessing the Web for public health surveillance », *The New England Journal* of *Medicine* vol. 360/21 (2009), 2153–2157.

Cioffi-Revilla, C. « Computational social science », in *Wiley Interdisciplinary Reviews : Computational Statistics*, 2/3 (2010), pp. 259-271.

Conte, R. et.al. « Manifesto of computational social science », in *European Physical Journal : Special Topics* 214 /1, 2012, pp. 325-346.

Driscoll, K. « From punched cards to 'Big Data' : A social history of database populism », in *Communication* +1/1 (2012), Online : http ://kevindriscoll.info/ 2012.

Bollier, D : The promise and peril of big data, Washington, DC : The Aspen Institute, 2012, Online : http ://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Eysenbach, G. « Infodemiology : Tracking Flu-Related Searches on the Web for Syndromic Surveillance », in : *AMIA Annual Symposium, Proceedings 8/2, (2006)*, 244-248.

Freyer-Dugas, A. et. al. « Google Flu Trends : Correlation With Emergency Department Influenza Rates and Crowding Metrics », in : *Clinical Infectious Diseases,* 54/15 2012, 463-469.

Fuchs, C. Social Media as Participatory Culture. In C. Fuchs, *Social Media : A Critical Introduction,* pp. 52-63. London and Thousand Oaks : SAGE 2014.

Dijck, J. van. *The Culture of Connectivity : A Critical History of Social Media*, Oxford : Oxford University Press 2013.

Galloway, A. (2004) : Protocol. How Control Exists after Decentralization. Cambridge/MA 2004.

Galloway, A. and Thacker, E. *The Exploit : A Theory of Networks*. University of Minnesota Press, 2007

Gitelman, L. and Pingree, G. B. *New Media : 1740-1915*, Cambridge, Mass. : MIT Press, 2004.

Ginsberg, J. et al. « Detecting influenza epidemics using search engine query data », in : *Nature. International weekly journal of science*, vol. 457 (2009), 1012-1014.

Golder, S. and Macy, M. « Social Science with Social Media ». In : *footnotes*, 40/1, 2012, Online : http ://www.asanet.org/footnotes/jan12/socialmedia_0112.html

Helmond, A. 2015. 'The Platformization of the Web : Making Web Data Platform Ready'. *Social Media + Society* 9, pp. 1–11.

Jordan. T. *Information Politics. Liberation and Exploitation in the Digital Society*. London : Pluto Press 2015.

Foucault, M. *The Order of Things : An Archeology of the Human Sciences.* New York : Pantheon Books 1970.

Lazer, D. et al. « Computational Social Science », in : *Science*, 323/5915, 2009, pp. : 721-723.

Lyon, D. Digital Citizenship and Surveillance | Surveillance Culture : Engagement, Exposure, and Ethics in Digital Modernity. *International Journal of Communication*,11. http ://ijoc.org/index.php/ijoc/article/view/5527 2017.

Mann, S., Nolan, J. & Wellman, B. (2003). Sousveillance : Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance & Society* vol. 1, no. 3, pp. 331-355. www.surveillance-and-society.org.

Manovich, L. "Trending : The promises and the challenges of Big Social Data", in : Matthew K. Gold (ed.), *Debates in the digital humanities*, Minneapolis : University of Minnesota Press 2012, pp. 460-475.

Manovich, L. « How to Follow Global Digital Cultures : Cultural Analytics for Beginners », in Becker, K. and Stalder, F., ed. *Deep Search : The Politics of Search beyond Google*, Edison, NJ, 2009, pp. 198–212.

Papacharissi, Z. (ed). *A Networked Self : Identity, Community, and Culture on Social Networking Sites*, London : Routledge 2010.

Paul, M. J. and Dredze, P.. « You Are What You Tweet : Analyzing Twitter for Public Health », in : *Proceedings* of the *Fifth International AAAI Conference* on *Weblogs* and *Social Media,* 2011, Online : www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/.../3264

Rogers, R. *Digital Methods*. Cambridge, Mass. : MIT Press 2013.

Trottier, D. & Fuchs, C. *Social media, Politics and the State : protests, revolutions, riots, crime and policing in the age of Facebook, Twitter and YouTube*. New York : Routledge 2015.

Walters, R. A. et.al. « Data sources for biosurveillance ». In : Voeller John G., ed. *Wiley handbook of science and technology for homeland security*, vol. 4. Hoboken : Wiley, 2010, 2431–2447.

6

TECHNIQUE, SOCIÉTÉ ET CYBERESPACE : LA GOUVERNEMENTALITÉ ALGORITHMIQUE

Marc Ménard et André Mondoux

Marc Ménard est professeur à l'École des médias et vice-doyen à la recherche et à la création de la Faculté de communication de l'UQAM. Ses travaux portent sur l'économie de la culture, les technologies numériques, les nouvelles formes de marchandisation de l'information et de la communication, et sur l'intelligence artificielle.

André Mondoux est professeur à l'École des médias de l'Université du Québec à Montréal (UQAM). Il s'intéresse aux liens entre technologies numériques et (re)production sociale autour de la banalisation de la surveillance, des circuits d'individuation psychique et collective liés au Big Data et des enjeux épistémologiques soulevés par l'industrialisation des médiations symboliques, l'Internet des objets et l'intelligence artificielle.

RÉSUMÉ

Après une phase triomphaliste, souvenons-nous des espoirs soulevés par « l'empowerment » des individus par les médias socionumériques, voici que ces derniers se retrouvent à nouveau à l'honneur, mais cette fois-ci en tant que source d'enjeux et de défis : sécurité et intégrité des données personnelles, propagation de fausses nouvelles, émergence des chambres à écho, ingérences étatiques dans les dynamiques électorales, etc. La plupart de ces problématiques, pour l'instant du moins, restent majoritairement confinées aux seuls domaines de la technique (posée comme agent d'optimisation neutre) et de l'individuel (comment assurer l'intégrité des données personnelles). Notre proposition vise à (ré) introduire le politique au sein de cet important débat, soit d'aborder la sécurité en termes de dynamiques sociales et sociétales. Pour ce faire, nous déploierons un cadre d'analyse inspiré de la notion de gouvernementalité algorithmique, c'est-à-dire porteuse et portée par des rapports de pouvoir induisant des formes de savoir et de subjectivité via une individuation économique, soit les circuits marchands des données personnelles, et qui peuvent avoir des conséquences importantes sur le plan du vivre-ensemble.

INTRODUCTION

Une des conséquences de l'utilisation de l'adjectif « cyber » est qu'il induit une prétendue séparation entre le monde (traditionnel) physique et le (nouveau) monde de « l'immatériel » et du « virtuel ». Pourtant, cette distinction, bien qu'intuitive, est sans cesse démentie : « être » dans le « cyberespace », par exemple, c'est aussi utiliser des ressources et infrastructures matérielles (ordinateurs, réseaux, etc.) en des endroits bien physiques ; participer à une économie politique (l'émergence du pouvoir des GAFAM[1]) et affronter des enjeux qui vont au-delà de la « culture » (le style de vie) pour s'immiscer dans les dynamiques sociopolitiques (intégrité des processus électoraux) et sociétales (industrialisation des médiations symboliques).

En un sens, le cyberespace s'affranchissait ainsi en quelque sorte du monde physique, un exercice de *tabula rasa* qui n'ouvrait rien de moins que les frontières d'un Nouveau Monde à construire et à habiter[2]. Et ce Cyber Far West s'est avéré radicalement différent du monde physique, principalement d'une part, par l'utilisation d'une technologie posée comme étant essentiellement neutre et, d'autre part, par la prédominance d'un individu – enfin – libéré de tout joug disciplinaire (« l'empowerment » du sujet hyperindividualiste).

Le « cyber » est le domaine par excellence de la technique livrée à elle-même. Posée comme neutre et pure rationalité instrumentale, la technique permettrait en ce sens de procéder à une médiation parfaite et transparente avec le réel ainsi abordé sans la « distorsion » de la représentation (la relativité propre au symbolique et sa nature idéologico-politique) : les données sont qualifiées de « brutes », la quantification permettrait d'atteindre une parfaite objectivité, les corrélations prédictives remplaceraient les théories et les relations de causalité, etc. En devenant pour ainsi dire sa propre médiation, la technique en vient à se poser en surplomb aux valeurs pour incarner l'objectivité des moyens sur les (relatives) finalités : « (…) la technique comporte comme donnée spécifique qu'elle se nécessite pour elle-même sa propre transformation. (…) C'est la

1. Google, Apple, Facebook, Amazon et Microsoft.
2. On se souviendra notamment de la *Déclaration d'indépendance du cyberespace* en 1996 par John Perry Barlow, un des fondateurs de *l'Electronic Frontier Foundation*, qui soutenait l'idée qu'aucune forme de pouvoir ne pouvait s'imposer et s'approprier l'Internet. https://www.eff.org/cyberspace-independence.

conjonction entre le phénomène technique et le progrès technique qui constitue le système technicien (Ellul, 2004 : 91). Royaume des moyens – et du « problem solving » – le cyber devient donc animé par des valeurs « non idéologiques », soit la concrétude de sa fonctionnalité, son efficience et son autoréférentialité : « Partout où il y a recherche et application de moyens nouveaux en fonction du critère d'efficacité, on peut dire qu'il y a technique » (Ellul, 2004 : 38).

À plusieurs égards, ce qui relevait du « système technicien » d'Ellul et du cyberespace est aujourd'hui considéré comme les caractéristiques générales des mondes – désormais unifiés – physique et « en ligne ». Non sans conséquences sur le vivre-ensemble, cette dynamique illustre bien ce que Freitag nommait le mode formel de reproduction de type décisionnel-opérationnel. Fondé sur la prémisse qu'une société se (re)produit dans le temps par le biais de son mode formel de reproduction symbolique, le mode de reproduction décisionnel-opérationnel atteste du déclin des médiations symboliques de type transcendantal[3] au profit des procédures opérationnelles objectifiées et transparentes avec les valeurs de pragmatisme, de fonctionnement et d'efficience technique :

> « C'est ainsi que la société consacre maintenant une activité incessante à sa propre unification. Elle est devenue une "société en autoproduction permanente", mais cette "autoproduction" n'est plus, comme dans le mode de reproduction politique classique, fondamentalement réflexive, mais "réactive". La société entre donc dans l'ère du système objectivé. Dans la constitution d'un tel système et la production de son unité de mode virtuellement purement objectif, la science joue alors un rôle essentiel, dans l'accomplissement duquel elle tend elle-même à accomplir sa propre mutation en technologisme et en technocratisme. (…) Ainsi, à la différence des sociétés primitives ou traditionnelles, les éléments de culture primaire qui sont drainés vers les mécanismes de la régulation décisionnelle ne comportent plus de leur côté aucune modalité d'intégration normative a priori, si ce n'est sur le plan de la procédure. (…). On assiste ainsi à un processus de totalisation sans totalité, et c'est précisément à cela que correspond le nouveau concept technocratique de "système" (…) » (Freitag, 1986 : 338-339).

Outre la technique, une autre caractéristique du monde « cyber » est qu'il est résolument de l'ordre de l'individuel, comme l'atteste notamment l'histoire des technologies numériques (Mondoux 2011a), des débuts des ordinateurs centraux aux technologies « individuelles » (ordinateurs

3. C'est-à-dire les valeurs, souvent institutionnalisées, qui font partie du *déjà-là* antécédent à l'émergence de l'individu et qui désigne donc les conditions de connaissance a priori.

personnels, tablettes, montres et téléphones intelligents, accessoires biométriques), jusqu'aux dispositifs liés aux dynamiques de personnalisation (moteurs de recherche, systèmes de recommandation, listes de lecture, etc.). Animé par la doctrine et les politiques du néolibéralisme, de même que par le déclin des grandes idéologies, le cyberespace est en effet devenu un monde à la mesure du libre arbitre individuel comme degré zéro de l'aventure humaine. C'est ainsi que les médias dits « sociaux » ont été traditionnellement abordés sur le plan de l'utilisation individuelle des technologies et que les enjeux soulevés, non sans surprise, relèvent également de l'ordre de la « vie privée » (intégrité/sécurité des données personnelles, liberté d'expression, etc.). Qui plus est, le fétichisme envers la technique (Dean : 2009) incite à voir celle-ci comme surgie *sui generis* (l'invention et la créativité du « 2.0 ») et que ce n'est qu'*après* que le social entre en scène en tant que simple accumulation d'intersubjectivités[4]. Plus qu'une simple tare individuelle (le soi-disant narcissisme inhérent aux médias sociaux), il s'agit d'une dynamique sociale en soi, soit celle de *l'hyperindividualisme* :

> « (…) L'émergence d'un sujet qui, réfutant ultimement toute forme d'autorité disciplinaire, aspire à advenir par et pour lui-même. Voilà pourquoi l'usage des médias socionumériques marque la nécessité, face au déclin des grands récits (Lyotard, 1979) et à l'idée que l'identité n'est plus ainsi "reçue" (construite au travers d'un collectif), de recourir à des stratégies communicationnelles afin de se construire une identité et l'affirmer aux autres » (Ménard et Mondoux, 2018 : 68-69).

Si je ne reçois plus mon identité de l'autre, je dois effectivement construire la mienne et la faire (re)connaître par les autres (Honneth, 2000), d'où la prolifération des jeux-questionnaires et sondages identitaires aux débuts des médias socionumériques en guise de pratiques d'auto-expression (Mondoux, 2011b).

En ce sens, hyperindividualisme et mode de reproduction décisionnel-opérationnel sont des plus complémentaires : le premier valorise sa propre jouissance (ici, maintenant) comme degré zéro du « social » (Melmann, 2002) et le second transforme toute valeur en simple opinion errante sur

4. C'est ainsi que le concept de *communauté*, à tort, a été déployé comme une version "à l'échelle" de la société : "Une communauté pure se conduirait comme un automate ; elle élabore un code de valeurs destiné à empêcher les changements de structure et à éviter la position des problèmes. Les sociétés au contraire, qui sont des groupements synergiques d'individus ont pour but de chercher à résoudre des problèmes" (Simondon, 2005).

une vaste mer de liberté individuelle. Si tous les individus sont égaux dans leur droit à la jouissance, se pose alors la question à savoir quelles valeurs – désormais réduites, libre arbitre oblige, au rang d'opinions tout égales en soi – pourraient légitimer une quelconque forme de régulation commune (sociale). C'est ici que la technique entre en jeu, en tant que médiation projetée comme neutre : c'est bien par la technique *immédiate* (sans médiation « idéologique » et en lien direct avec le « réel ») que la technique permet au sujet d'être vraiment « lui-même », c'est-à-dire émancipé de l'idéologique et du politique (le fameux *empowerment*) dans la durée de *l'immédiat*, soit le temps réel de la vélocité technique et de la jouissance (Ménard et Mondoux, 2018).

Afin de bien saisir les liens entre hyperindividualisme et hégémonie technique, caractéristiques des dynamiques de type « cyber », il est nécessaire de pousser la réflexion au-delà des dualités entre sujet et société, technique et politique. Pour ce faire nous mobiliserons le concept foucaldien de gouvernementalité (Foucault, 2012), plus précisément une inspiration contemporaine désignée sous le nom de *gouvernementalité algorithmique* (Rouvroy et Berns, 2013).

GOUVERNEMENTALITÉ ALGORITHMIQUE

Le concept de gouvernementalité chez Foucault déploie les notions phares du philosophe, soit discours, pouvoir, vérité et subjectivité ; il exige :

> « (…) Que l'on place au centre de l'analyse non le principe général de la loi, ni le mythe du pouvoir, mais les pratiques complexes et multiples de gouvernementalité qui suppose d'un côté des formes rationnelles, des procédures techniques, des instrumentations à travers lesquelles elle s'exerce et, d'autre part, des enjeux stratégiques qui rendent instables et réversibles les relations de pouvoir qu'elles doivent assurer » (Foucault, 1994 : 584).

Ainsi, pour Foucault, la gouvernementalité :

> « (…) C'est se donner les moyens de mieux comprendre les modalités par lesquelles l'action publique s'efforce d'orienter les relations entre la société politique (via l'exécutif administratif) et la société civile (via ses sujets administrés), mais aussi entre les sujets eux-mêmes. » (Lascoumes, 2004).

Il s'agit d'une forme de régulation générale axée vers la *production et circulation structurées des énoncés et discours* :

> « Dans toute société la production de discours est à la fois contrôlée, sélectionnée, organisée et redistribuée par un certain nombre de procédures qui ont pour rôle d'en conjurer les pouvoirs et les dangers, d'en maîtriser l'événement aléatoire, d'en esquiver la lourde, la redoutable matérialité. » (Foucault, 1971 : 11).

D'une part, ceci induit un *régime de vérité*, c'est-à-dire ce qui est considéré comme vrai ou faux et qui délimite ainsi les champs du possible :

> « Ce n'est pas l'activité du sujet de connaissance qui produirait un savoir, utile ou rétif au pouvoir, mais le pouvoir-savoir, les processus et les luttes qui le traversent et dont il est constitué, qui déterminent les formes et les domaines possibles de la connaissance » (Foucault, 2015 : 289).

D'autre part, est également induit le rapport de pouvoir créé par l'acceptation par les sujets de cette vérité et qui du coup se trouve à individuer non seulement le rapport de pouvoir (re)produire les discours, mais également les modalités d'assujettissement, c'est-à-dire la production du sujet lui-même :

> « En fait, ce qui définit une relation de pouvoir, c'est un mode d'action qui n'agit pas directement et immédiatement sur les autres, mais qui agit sur leur propre action. Une action sur l'action, sur des actions éventuelles ou actuelles, futures ou présentes » (Foucault, 2001 : 1055)

Et c'est par ce retour du pouvoir vers le sujet, que la boucle de la gouvernementalité est complétée :

> « (…) Dans ce jeu la liberté va bien apparaître comme condition d'existence du pouvoir (à la fois son préalable, puisqu'il faut qu'il y ait de la liberté pour que le pouvoir s'exerce et aussi son support permanent puisque, si elle se dérobait entièrement au pouvoir qui s'exerce sur elle, celui-ci disparaîtrait du fait même et devrait se trouver un substitut dans la coercition pure et simple de la violence) (…) » (Foucault, 2001 : 1057).

Adaptée au contexte du numérique, la gouvernementalité devient *algorithmique* et renvoie alors à « un certain type de rationalité (a) normative ou (a) politique reposant sur la récolte, l'agrégation et l'analyse automatisée de données en quantité massive de manière à modéliser, anticiper et affecter par avance les comportements possibles » (Rouvroy et Berns, 2013).

À bien des égards, la notion de gouvernementalité algorithmique demeure essentiellement de l'ordre de l'abstraction, c'est-à-dire qu'elle doit être dégagée et reconstruite par l'analyse. Afin de mieux saisir les principaux éléments de la gouvernementalité algorithmique, nous invoquerons deux notions de Simondon : la *transduction* et le *milieu associé*.

Simondon définit la transduction comme étant « une opération physique, biologique, mentale, sociale, par laquelle une activité se propage de proche en proche à l'intérieur d'un domaine, en fondant cette propagation sur une structuration du domaine opérée de place en place » (Simondon, 2005 : 32). Autrement dit, la transduction désigne l'émergence d'un rapport (dit *transductif*) produit par la mise en relation d'éléments, rapport qui en retour (re)produit ces éléments individuellement. La gouvernementalité algorithmique est transductive dans la mesure où, en tant que rapport, elle n'implique pas l'intervention de facteurs extérieurs à ses éléments constitutifs, que ce soit une vision substantialiste prédéterminée du sujet (l'Homme), la notion d'un pouvoir pouvant être détenu ou toute « pré nécessité » épistémologique (la Vérité).

La deuxième notion, celle du milieu associé, n'est pas sans lien avec la pensée heideggérienne : ce qui *est*, est ce qui se (re)produit dans le temps, et ce, dans l'incontournable espace du monde (la mondanité comme totalité des places et le sujet heideggérien – *Dasein* – comme « l'être-au-monde »). Ainsi, le sujet arrive toujours dans une réalité *déjà-là* (Heidegger) ou *pré-individuelle* (Simondon). Pour Simondon, il s'agit du milieu associé ; associé parce qu'il est co-instituant dans la mesure où le processus d'individuation (apparition dans le monde) est une récurrence de causalité avec un milieu associé :

> « L'individu serait alors saisi comme une réalité relative, une certaine phase de l'être qui suppose avant elle une réalité pré-individuelle, et qui, même après l'individuation, n'existe pas toute seule, car l'individuation n'épuise pas d'un seul coup les potentiels de la réalité pré- individuelle et d'autre part, ce que l'individuation fait apparaître n'est pas seulement l'individu, mais le couple individu-milieu » (Simondon, 2005 : 24-25).

L'approche de l'individualisme méthodologique (poser le sujet comme degré zéro du social), n'éclaire en rien la genèse de ce qui est un « sujet », c'est-à-dire un individu producteur, certes, mais également lié à un processus d'assujettissement. Sous cet angle, l'individuation, qu'elle soit biologique, psychique ou collective, n'est qu'une phase du processus d'individuation et l'individu est ainsi toujours « en devenir », il est perpétuellement « à être » ou, pour reprendre le terme heideggérien, il est un « être déjà-en-avant-de-soi-dans-le-monde ». C'est ainsi que nous mobiliserons la *circulation marchande des données personnelles* à la fois comme réalité pré-individuelle/déjà-là (les structures économiques et politiques en place au moment de l'émergence du « cyber ») et milieu associé (l'économie comme facteur co-instituant dans l'individuation du « cyber »). Nous entendons donc

démontrer comment la notion de gouvernementalité algorithmique, telle que dégagée par la circulation marchande des données, permet de saisir de quelle façon le sujet produit des rapports sociaux qui à leur tour produisent un sujet à (re)produire ces rapports sociaux, et quel est l'apport et les conséquences de la technique (le numérique) à cet égard.

CIRCULATION MARCHANDE ET GOUVERNEMENTALITÉ ALGORITHMIQUE

Débutant par le sujet *productif*, tout en gardant à l'esprit que celui-ci sera à son tour *produit* (rapport de transduction), la circulation marchande des données personnelles est constituée de plusieurs moments-clés : la production de données par le sujet ; la captation des données ; leur stockage et analyse ; et le retour communicationnel vers le sujet effectuant le bouclage du circuit :

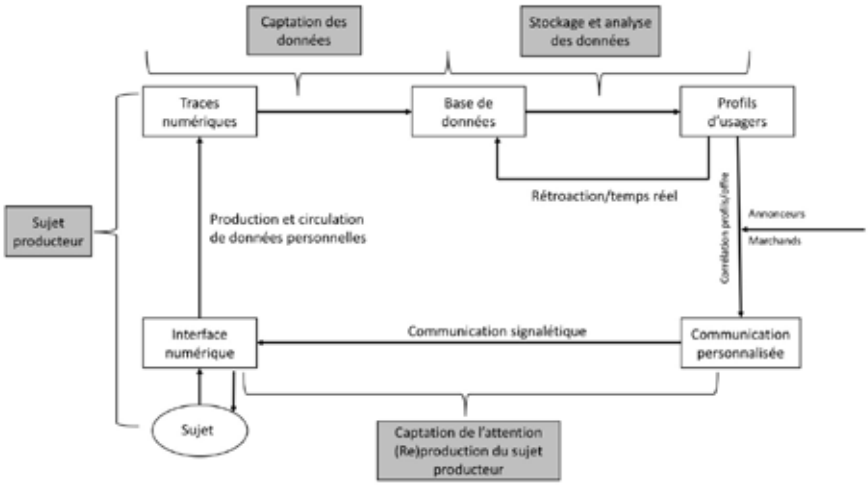


Figure 1 : Circulation marchande des données personnelles

(Ménard et Mondoux, 2018, p.67)

Une des caractéristiques des technologies numériques est qu'elles sont mnémoniques (Stiegler, 1994) et en ce sens capables de conserver toutes traces de leur utilisation. Le sujet productif, en mobilisant l'outil technique (hyperindividualisme et pratiques auto-identitaires), génère ainsi des

traces : les données personnelles. D'emblée, dès le départ et à l'encontre de l'idée reçue que la technique serait neutre, il faut souligner le rôle de médiation exercé par celle-ci. En effet, qu'il s'agisse de rédiger avec 240 caractères d'espace ou trois pages, en deux heures ou en quelques secondes, les caractéristiques de l'outil lui confèrent un rôle de codétermination des énoncés produits comme « vérité » du sujet :

> « Dans une perspective de gouvernementalité, cette "vérité" reste tributaire de son individuation sociohistorique, c'est-à-dire de rapports de pouvoir qui marquent à la fois la normativité (l'ordre, le sens) et l'expressivité (le politique, la relativité du signe). Ces rapports sont actualisés à même l'interface-usager : ainsi le formatage binaire des données limite l'horizon de possibilités (oui/non), les "choix" sont préformatés et par-dessus tout, ce ne sont pas toutes les expressions possibles qui sont effectivement numérisées et captées. » (Ménard et Mondoux, 2018 : 71)

De plus, il faut également souligner que l'horizon symbolique du sujet est radicalement réduit sous forme de *comportement*, soit l'action de cliquer sur une icône (« J'aime » par exemple pour Facebook). C'est ainsi que cliquer pour « aimer » une montre antique, par exemple, ne révèle en rien la richesse symbolique derrière celle-ci : on peut aimer l'objet en tant que montre, antiquité, évocation d'un parent, aubaine commerciale, etc. Comme on le constate, la « vérité » des énoncés du sujet se fait déjà des plus malléables… Ce sont ces traces, sous forme d'abstraction, qui sont colligées.

En utilisant l'objet technique, le sujet se trouve ainsi à actualiser les rapports de pouvoir, notamment lors de la phase de captation des données personnelles qui sont déjà, à ce stade, étroitement paramétrées par les interfaces en fonction des besoins des fournisseurs de services. En effet, la captation de ces données, particulièrement en contexte du temps réel comme idéal, d'une synchronicité parfaite avec le réel (et temporalité de la jouissance du sujet), entraîne un monitorage continu, soit une forme de *surveillance persistante* à laquelle les sujets s'y prêtent, volontairement ou non (Royvroy et Berns, 2013). De plus, autre dimension des rapports de pouvoir, les contrats d'utilisation de la plupart des services socionumériques (Facebook, Twitter, Instagram, etc.) permettent au fournisseur de partager ces données avec de tierces parties sans pour autant assumer la responsabilité de ce que celles-ci feront avec les données. Autrement dit, le sujet producteur ne peut rien dire ou faire sur le sort de ses données personnelles et est exclu de leur processus de valorisation économique.

C'est lors de la phase de stockage et d'analyse qu'interviennent les dynamiques du Big Data et de l'intelligence artificielle : création de bases de données, analyses statistiques, corrélations prédictives et création de profils individuels, le tout automatisé et fonctionnant en temps réel. Sur le plan épistémologique, cette phase se caractérise par sa prétention, décontextualisation et désaffection des données et neutralité supposée de la technique aidant, à représenter le « réel » sans médiation : les *données brutes*. Quantifiées et mathématiquement manipulables, les données brutes sont posées comme objectives et constituant du « réel » et à ce titre, considérées *vraies*. Ces données sont perçues comme étant d'autant plus vraies qu'elles sont produites par, pour et sur le sujet lui-même (et en cela révélateur de l'hyperindividualisme tel que décrit précédemment). C'est ainsi que le « réel » peut désormais être accédé sans l'apport d'une extériorité, que ce soit le symbolique (la représentation), le politique (l'autre), la science (les pairs), comme n'importe quelle autre ressource « naturelle ». Conformément à la pensée foucaldienne, on constate ici que l'ordre du discours consiste effectivement à structurer la production et la circulation des énoncés tout en actualisant un régime de vérité, soit un ensemble de croyances tenues pour vraies et qui ainsi font que le sujet, en adhérant à celles-ci, participe à une dynamique d'assujettissement qui se fait simultanément *soumission* et *production du sujet*.

Au sein de la circulation marchande, le profil de l'usager, créé par captation et analyse des données personnelles du sujet, joue un rôle central dans le processus d'assujettissement. À cet effet, la notion de *double numérique* est particulièrement révélatrice, mais trompeuse, car elle suggère une adéquation naturelle entre les données brutes et le sujet lui-même. En effet, le profil masque ainsi le travail de décontextualisation et d'abstraction des énoncés symboliques du sujet, de même que les stratégies des acteurs économiques. En effet, le profil n'est pas identité, mais plutôt *identification*, entendu ici comme le résultat d'une action visant à identifier un individu d'après un ensemble de critères préétablis répondant aux objectifs des fournisseurs de services (Bonenfant *et al.* 2015). Il s'agit en effet, à cette phase, de créer des profils d'usager en assemblant les données « brutes » contenues dans une base de données. L'objectif d'un profil est de représenter les intérêts, caractéristiques et préférences spécifiques d'un usager qui seront « susceptibles d'aider le système d'information à fournir les données les plus pertinentes, dans la bonne forme, au bon endroit et au bon moment » (Bouzeghoub et Kostadinov, 2006 : 2), c'est-à-dire de convoquer l'usager dans l'immédiateté de sa jouissance. Sous cet angle se manifeste

la production du sujet hyperindividualiste : le double numérique serait la version parfaite et objective du sujet ; d'un sujet qui pourrait se regarder avec le miroir neutre et objectivant de la technique exposant ses données personnelles ; une révélation de soi par soi-même.

LA PRODUCTION DU SUJET

L'assujettissement est également induit par le processus d'individuation de la gouvernementalité, c'est-à-dire les rapports de pouvoir et le régime de vérité. Ainsi, la notion de double numérique renvoie à une conception de l'identité où celle-ci n'est plus une construction sociale impliquant l'autre, que ce soit le stade du miroir où l'enfant apprend par le regard des autres à prendre conscience de son corps et à le distinguer des autres corps (Lacan, 1949), ou la reconnaissance de cette identité par les autres (Honneth, 2000). Au contraire, le double numérique se présente comme l'irruption du réel, une vérité objective qui est synthétisée, très souvent, sous forme de tableau de bord (*dashboard*) permettant la visualisation des données brutes. La plupart du temps, ces données sont de l'ordre du quantitatif et traitées statistiquement, ce qui permet non seulement de visualiser et synthétiser les activités du sujet, mais également d'effectuer des projections et tendances pour ainsi servir d'indicateurs de progression et de performance. Comme on peut le constater, cette vision identitaire est conforme au mode décisionnel-opérationnel de reproduction sociale chez Freitag caractérisé par les valeurs de pragmatisme, d'optimisation et d'efficience.

Le processus d'assujettissement comporte également une part active du sujet lui-même. De par son adhésion au régime de vérité (ici, les données brutes quantitatives sont objectives et le double numérique le représente), le sujet en vient à actualiser et à (re)produire le régime de vérité par autodiscipline, induisant du coup, le rapport de pouvoir qui se manifeste par « l'action sur des actions futures » de Foucault. Ainsi, prenons l'exemple d'un tableau de bord qui indique, par exemple, que vous avez marché 4 000 pas aujourd'hui. Voilà une information. Si le tableau de bord indique que cela représente 500 pas de plus que la veille pour un total de 48 % de l'objectif quotidien à atteindre, il s'agit également d'une information (mise en forme) de la subjectivité en incitant le sujet à marcher davantage et ainsi être plus performant. De plus, cette dynamique incite le sujet à produire davantage de données brutes que ce soit pour augmenter ses performances, préciser et peaufiner ses préférences personnelles ou partager ses résultats sur les médias socionumériques par le biais de l'utilisation de l'outil technique

pour (re)produire les stratégies d'expression identitaire propres au sujet hyperindividualiste et ainsi boucler le circuit de la gouvernementalité sur lui-même (transduction).

Sur le plan de l'économie politique, notons que cette dynamique est tout à fait l'avantage des fournisseurs de services dont le modèle d'affaires, qu'il s'agisse des médias socionumériques ou des entreprises de Big Data et d'intelligence artificielle, est justement la nécessité de colliger et d'analyser des quantités massives de données afin d'en effectuer la valorisation. En ce sens, la notion contemporaine de *présumer* est révélatrice : voici le sujet simultanément producteur et consommateur.

Une autre caractéristique du processus d'assujettissement au sein de la gouvernementalité algorithmique est le régime temporel – le *temps réel* – dans lequel il se déploie. Le temps réel, en effet, correspond à la fois à la vélocité idéale de la technique (Kitchin et McArdle, 2016) et au moment de la jouissance chez le sujet (Mondoux, 2011b). Voilà pourquoi plusieurs auteurs parlent d'économie de l'attention (Kessous, 2012) ou d'économie de la pulsion (Stiegler, 2015). Ce régime temporel tend à transformer la communication en signalétique (par exemple, rouge = « cessez de marcher » et vert = « marcher davantage »), c'est-à-dire en sémiotiques a-signifiantes : Le « message » ne passe pas par des chaînes linguistiques, mais par le corps, des postures, des bruits, des images, des mimiques, des intensités, des mouvements, des rythmes, etc. » (Lazarrato, 2006). Ces sémiotiques sont *machiniques* dans la mesure où 1) elles s'adressent et opèrent de façon intuitive et impulsive, 2) elles intègrent l'individu dans un circuit général où le sujet agit à titre de relais automatique (temps réel) afin de produire et faire circuler les données nécessaires aux circuits marchands et 3) elles font la reproduction même en confinant l'horizon des possibles du sujet aux prévisions de ses actions faites à partir de ses caractéristiques et comportements passés.

RETOUR SUR LE CYBER, LE SUJET ET LA SÉCURITÉ

La notion de gouvernementalité algorithmique permet non seulement d'éclairer les rapports entre technique, individu et société, mais également de réfuter certaines conceptions qui alimentent aujourd'hui les réflexions autour de la technique telle que déployant le « cybermonde ». Ainsi la vision d'un sujet comme degré zéro du réel est irréaliste : à bien des égards le sujet émerge en rapport avec des dynamiques sociétales et de ce fait, comme l'a

toujours indiqué Simondon, les individuations psychiques et collectives sont étroitement reliées. Sous cet angle, le sujet est effectivement une phase de son processus d'individuation. Deuxième conception à corriger à propos du sujet : il est de plus en plus difficile de le voir conforme au modèle économique traditionnel, c'est-à-dire comme un être de raison effectuant constamment des choix rationnels censés maximiser son utilité individuelle. La gouvernementalité algorithmique suggère au contraire que la vélocité technique, couplée avec les dynamiques de jouissance crée des rapports relevant davantage du conditionnement des sujets confinés dans des boucles sans fin de jouissance où la satisfaction des désirs produit un vide sans cesse à combler (Dean, 2009) et où la rationalité de la prise de décision du sujet est en partie déléguée à un dispositif technique.

La notion d'une technique neutre est également à nuancer. En apparaissant ainsi, la technique ouvre toutes grandes les portes d'un « réel » enfin harnaché sans les défis du politique et de la représentation. Un réel qui serait pure rationalité instrumentale. Le principal problème avec cette approche, c'est que la technique risque de devenir le but en soi (et pour soi) et :

> « Cela signifie céder à une exigence qui se situe au-dessus de l'homme, au-dessus de ses projets et de ses activités. Ce que la technique moderne a d'essentiel n'est pas une fabrication purement humaine. L'homme actuel est lui-même provoqué par l'exigence de provoquer la nature à la mobilisation. L'homme lui-même est sommé, il est soumis à l'exigence de correspondre à ladite exigence » (Heidegger, 1990 : 30).

Voilà pour Heidegger, qui déjà à l'époque, lançait un cri d'alarme à propos de notre utilisation de la technique : « Peut-être est-il une pensée plus sobre que le déferlement irrépressible de la rationalisation et l'emportement qu'est la cybernétique. C'est plutôt cet emportement qui pourrait bien être le comble de l'irrationnel » (Heidegger, 1990 : 138). Pour Heidegger, le danger réside à réduire l'humain à un simple statut de mise en disponibilité pour les exigences de la technique, tout en nourrissant l'illusion de sa puissance et de sa totale maîtrise de celle-ci.

Un autre défi est que le mode de régulation associé à la gouvernementalité algorithmique tient davantage de l'ordre du *contrôle*, une valeur phare de la cybernétique (Wiener, 1948). D'une part, parce que la dynamique d'assujettissement tend à confiner le sujet dans des rapports de conditionnement. Ceci n'est guère surprenant lorsqu'on connaît les liens étroits entre la cybernétique et le béhaviorisme (Teixeira, 2015). D'autre

part, en étant confronté au « réel », le sujet ne peut en rien « négocier » : il est ou pas conforme au réel. Il s'agit donc moins de discipline et faire de son mieux pour endosser et intégrer les valeurs tout en conservant une marge politique (il y a des valeurs avec lesquelles je suis d'accord ou non) – que de contrôle : le réel ne souffre d'aucun compromis. Sur le plan de vivre ensemble, cela se traduit par la priorité des *faits* sur la loi ; les lois relevant du politique et donc négociables, tandis que les faits ne sont pas discutables. Nous nous retrouvons devant une contradiction fondamentale de la gouvernementalité algorithmique : sur le plan individuel, tout est permis (n'importe quoi), mais collectivement nous serions mus par des dynamiques d'efficience (tout sauf n'importe quoi).

Ceci marque le déclin du politique où la notion de *praxis*, soit l'ensemble des activités vouées à la prise en charge du destin collectif, devient désuète, occultée par la technique qui tend à la subsumer. À cet effet, il n'y a qu'à constater comment il n'est plus nécessaire d'obtenir la participation directe et volontaire du sujet dans les formes de « synthèse collectivité » produites avec l'aide de la technique, que ce soit l'intelligence collective, le Big Data ou l'extrapolation des valeurs politiques des individus via leurs comportements en matière de consommation. Ainsi, le temps et l'espace pour la réflexion collective tendent à s'estomper sous le poids de l'hyperindividualisme, du temps réel, des dynamiques de conditionnement et du contrôle comme de la régulation (Mondoux *et al.* 2016).

Ce déclin du politique, a été, est et est encore observé par plusieurs auteurs sous plusieurs formes, que ce soit *l'effondrement symbolique* (Bougnoux, 2006), la *perte d'efficience symbolique* (Žižek, 2009) ou la *misère symbolique* (Stiegler, 2013). C'est cette dimension, pratiquement absente des débats gravitant autour de la sécurité dans le cyberespace (ou de l'intelligence artificielle), que réside une menace ainsi ignorée et qui selon nous devrait être incluse dans les éléments à *sécuriser* lorsque l'on parle de cybersécurité.

Sur le plan de la politique, le déclin des institutions (par essence de nature transcendantale), pave la voie à l'émancipation des extrémismes à tendance totalitaire qui un peu partout sur la planète (re)font leur apparition.

> « L'idée totalitaire de monde administré, dans lequel l'expérience même de la liberté subjective est la forme apparente de la sujétion aux mécanismes disciplinaires, n'est en dernier ressort que l'envers fantasmatique et obscène de l'idéologie (et de la pratique) publique officielle, celle de l'autonomie et de la liberté individuelle » (Žižek, 2009 : 144.).

Et plus fondamentalement encore, sur le plan *du* politique :

> « (…) ce sont ni plus ni moins les fondements ontologiques, anthropologiques et écologiques de "l'être ensemble" qui sont ainsi menacés, non seulement parce que le "système" tend à imposer sa logique technocratique et économique au "monde vécu" (Habermas), mais parce que le monde vécu lui-même se trouve de plus en plus éclaté et éparpillé par les effets désocialisant d'un hyperindividualisme qui se nourrit d'une conception abstraite et formelle de la liberté (Bischoff, 2008 : 152-153).

Il serait facile de balayer du revers de la main cette position comme étant fallacieusement technophobe. En effet, en cette période d'investissements massifs en matière de technologies numériques, d'intelligence artificielle et de cybersécurité, tout ce qui ne va pas dans le sens (pré)convenu passe pour son contraire. Il serait évidemment de mauvaise foi de nier tout avantage et bienfait liés au déploiement des technologies numériques. Il le serait tout autant de nier que les enjeux de sécurité, c'est-à-dire s'assurer de l'absence de tout danger, se limiteraient aux seules dimensions d'une individualité ivre de sa liberté et d'une technique neutre et entièrement maîtrisée. Comme le souligne Dubois "en ce sens, l'époque de la technique pourrait bien être le règne du 'sans question', l'évidence équivoque d'une fonctionnalité parfaite, d'où la maîtrise humaine de la nature jouerait le leurre par excellence" (Dubois, 2000 : 138).

BIBLIOGRAPHIE

Bischoff, M. (2008), "Une brève présentation de la sociologie dialectique de Michel Freitag", *Économie et Solidarités*, 39 (2).

Bonenfant, M., M. Ménard, A. Mondoux et M. Ouellet (2015), "De l'identité à l'identification. La dérive du tiers symbolisant", in Bonenfant, M. et C. Perraton dir., *Identité et multiplicité en ligne*, coll. Cahiers du Gerse, Presses de l'Université du Québec, Sainte-Foy, p. 25-49.

Bougnoux, D. (2006), *La crise de la représentation*, Paris, Éditions La Découverte.

Bouzeghoub M. et D. Kostadinov (2006), "Data Personalization : a Taxonomy of User Profiles Knowledge and a Profile Management Tool, CNRS : UMR8144 – Université de Versailles Saint-Quentin-en-Yvelines. http ://hal.archives-ouvertes.fr/hal-00141372.

Dean, J. (2009), *Democracy and Other Neoliberal Fantasies*, Duke Press.

Dubois, C. (2000), *Heidegger, Introduction à une lecture*, Paris, Seuil.

Ellul, J. (2004) [1977], *Le Système technicien*, Paris, Le cherche midi.

Foucault, M. (1971), *L'ordre du discours*, Paris, Gallimard.

Foucault, M. (2015) [1975], *Surveiller et punir. Naissance de la prison*.

Foucault, M. (2001) [1984], 'Préface à *L'histoire de la sexualité*', *Dits et écrits*, T. IV, Paris, Gallimard.

Foucault, M. (1994) [1984], texte de 1984, *Dits et écrits*, T. IV, 1994.

Freitag, M. (1986), *Dialectique et société*, Tome 2, Montréal, Saint-Martin.

Heidegger, M. (1990), *Langue de tradition et langue technique*, Lebeer-Hossmann.

Honneth, A. (2000), *La Lutte pour la reconnaissance*, Paris, Cerf.

Kessous, E. (2012), *L'attention au monde : Sociologie des données personnelles à l'ère numérique*, Paris, Éditions Armand Colin.

Kitchin, R., et McArdle, G. (2016), 'What makes Big Data, Big Data ? Exploring the ontological characteristics of 26 datasets", *Big Data & Society*. https://doi.org/10.1177/2053951716631130

Lacan, J. (1949), 'Le Stade du miroir comme formateur de la fonction du Je : telle qu'elle nous est révélée dans l'expérience psychanalytique', *Revue française de psychanalyse*, p. 449-455.

Lascoumes, P. (2004), 'La Gouvernementalité : de la critique de l'État aux technologies du pouvoir', Le Portique [En ligne], 13-14 | 2004, mis en ligne le 15 juin 2007, consulté le 21 juin 2019. URL : http://journals.openedition.org/leportique/625.

Lazzarato, M. (2006), Le 'pluralisme sémiotique' et le nouveau gouvernement des signes. Hommage à Félix Guattari' in Transversal - eipcp multilingual webjournal à http://eipcp.net/transversal/0107/lazzarato/fr le 28 juin 2019.

Lyotard, J.-F. (1979), *La condition postmoderne*, Paris, Les éditions de minuit.

Melmann, C. (2002), *L'homme sans gravité*, Paris, Gallimard.

Ménard, M. et Mondoux, A. (2018), "Big Data, circuits marchands et accélération sociale" in *Big Data et société* (Mondoux, Ménard éd.), PUQ, p.p. 68-69)

Mondoux, A., Ménard, M., Bonenfant, M. et Ouellet M. (2016), "Big Data et quantification de soi. La gouvernementalité algorithmique dans le monde numériquement administré ', *Canadian Journal of Communication*. 40 (4) : 597-613.

Mondoux, A. (2011 a), *Histoire sociale des technologies numériques*, Montréal, Nota Bene.

Mondoux, A. (2011b), "Identité numérique et surveillance", *Les Cahiers du numérique, L'identité numérique* - Volume 7, N° 1/2011, 49-59, Paris, Éditions Lavoisier, août 2011.

Rouvroy, A. et Berns, T. (2013), "Gouvernementalité algorithmique et perspectives d'émancipation : Le disparate comme condition d'individuation par la relation ?", *Réseaux*, 177 (1), pp. 163-196.

Simondon, G. (2005), *L'Individuation à la lumière des notions de forme et d'information*, Paris, Jérôme Million.

Stiegler, B. (1994), *La Technique et le temps*, tome 1 : *La Faute d'Épiméthée*, Paris : Éditions Galilée.

Stiegler, B. (2013), *De la misère symbolique*, Paris, Flammarion.

Stiegler, B. (2015), *La société automatique,* Paris, Galilée.

Teixeira, A. (2015), "The Pigeon in the Machine. The Concept of Control in Behaviorism and Cybernetics', Pasquinelli, M. (ed.), *Alleys of Your Mind. Augmented Intelligence and Its Traumas*, Lüneburg : meson press, pp. 23-34.

Wiener, N. (1948), *Cybernetics : Or Control and Communication in the Animal and the Machine,* Mass. MIT Press.

Žižek, S. (2009), *Bienvenue dans le désert du réel*, Paris, Champs essais.

7

THE CASE OF FAKE NEWS AND AUTOMATIC CONTENT GENERATION IN THE ERA OF BIG DATA AND MACHINE LEARNING

Nicolas Garneau

Nicolas is currently a Ph.D. candidate at Laval University, Québec, within the Group of Research in Artificial Intelligence of Laval. His research focuses on the natural generation of textual content for less represented languages using deep learning models. He specializes in transfer learning and few-shot learning. Nicolas has several years of experience in software development and data analysis. Having recently taught the deep learning class at Laval University, he is an excellent speaker and he enjoys giving conferences and talks about artificial intelligence. He is also a founding member of the Baseline Cooperative, a group of researchers that seek to democratize artificial intelligence by supporting companies in their everyday business processes.

INTRODUCTION

Lately popularized by the United States of America's president during the 2016 United States presidential elections, Donald Trump, we have seen an impressive ascent of the term "fake news" in the media all over the world. While Trump used this term to uncover falsified claims reported by journalists, the concept of fake news is broad and extends far beyond defamation.

Fake news dates way before these United States presidential election and has been used by the human race since the beginning of our civilization [1]. Fake news can, unfortunately, be seen as a powerful weapon. In the context of elections, it can be used to taint a person's reputation, but it can also be used as propaganda. Indeed, it has been widely used in several communities and countries to convey an ideology forcing people to think in some way or believe in a specific thing that would benefit the actor behind the propaganda. The conveyed information is not impartial and is used to influence the audience's mind about a particular movement. The first major form of large-scale propaganda has been seen in the early days of the Great War in 1914 where governments encouraged their respective citizens to get enrolled in the army to fight for their country's interest.

Another form of fake news is simply pure unintended misinformation. One can mistakenly report facts with neither a good nor bad intention. If this fake news goes viral and can potentially harm a person's reputation, it will usually be debunked rapidly.

Humorous fake news has inevitably taken over social media. We see websites dedicated to specially craft humorous fake news about daily events in order to build and maintain a community of readers, hence making money. Even though, this method is not intended to hurt anybody, it too often crosses that fuzzy line between humour and defamation.

This brief list of different types of fake news [2] has shared a similar key component to date : A human is behind the generation of such fake news. What if this generation could be automated ? We actually drown in information with the countless media sources that present human-generated and curated content. Just the idea of automatically generated content can make us dizzy. Unfortunately, with the rise of Artificial Intelligence and the vast amount of data available on the web (i.e. the Big

Data era), we are much closer to this inevitable drowning than we can think of.

In the next section, we will uncover the recent advances in artificial content generation such as natural language [19] and video [20] that have seen the day due to novel Machine Learning architectures, especially Deep Neural Networks [3]. We will precisely analyze how an Artificial Intelligence model can learn to generate fake content, also called DeepFakes, based on a large amount of data. We will also outline what are the needs in order to achieve imperceptible generation from the human eyes. As with any domain where artificial intelligence is applied, we can see machine learning models for fake content generation as a facilitator where the user can easily generate falsified content for later fine-tuning, as we will see in Section 3.2.

For each way of generating falsified content, we will present preliminary research that has been conducted in order to fight counterfeit generated content and how can we possibly fight this emerging phenomenon. We will conclude this chapter with the possible future directions on artificially generated content.

AUTOMATIC GENERATION OF FAKE CONTENT

As previously stated, the generation of fake content (i.e. fake news) is actually mainly initiated by a human. However, with the rise of machine learning and the huge amount of data available on the internet, it is likely that artificial intelligence will be (or is actually used!) to automate the generation fake content or at least facilitate the automation.

In this section, we will go through three different types of content that can be generated by an artificial algorithm : visual content, audio and textual natural language. For each medium, we will analyze which type of machine learning model can be used to generate that specific kind of content as well as the data needed to train such architectures.

AUTOMATIC GENERATION OF VISUAL CONTENT

The automatic generation of visual content has received much attention with the rise of deep learning and the Generative Adversarial (GAN) architecture proposed by Goodfellow et al. A GAN is inspired by Game Theory where a generator generates a piece of information (e.g. an image)

and a discriminator tries to distinguish a real from a generated piece of information. This algorithm has been trained with images where the generator learns to fool the discriminator by generating realistic images.

The graphic card manufacturer giant, NVIDIA, showed in 2017 that it was possible to generate fake images of celebrities using GANs [9]. Their architecture was trained on 30,000 images of size 1024 per 1024 using 8 Tesla V100 graphic cards in parallel for four days. While this training configuration seems affordable [4], one does not come up with this architecture out of the blue. The total cost of obtaining such a model is much more expensive. A handful of generated images can be seen in the original paper [9]. It is interesting to see that, while some images are really impressive, the generation is far from perfect in many cases.

Recently, a group of researchers extended the fake generation of content to videos [20]. Indeed, by using a special type of recurrent neural network, a Long Short-Term Memory (LSTM) network [7], they were able to produce artificial video sequences of the United States president, Barack Obama. Their model learns to align an audio segment of the president with a high-quality video of him speaking. One may recall the campaign conducted in order to highlight the increasing sophistication of DeepFakes where Jordan Peel is literally putting words in the mouth of the president. Words that he, we assume, would not say in a public address. The video [5] has aroused even more the interest towards the fake generation of content.

Even though we can perceive generation artifacts when we look closely at the images, we can assume that the deep learning models will just keep getting better at generating fake content. This is why many researchers [16, 6, 1, 12], just to name a few] are studying the question ; are we able to detect falsified generated visual content ? The main idea that gets out of the current results is that a model that is really good at generating fake images or videos will be at least equally good at detecting them. Facebook [6] and Google [7] also put a lot of resources in order to stimulate the community to develop robust solutions in order to detect DeepFakes.

AUTOMATIC GENERATION OF TEXTUAL NATURAL LANGUAGE

The whole community of Natural Language Processing (NLP) has been taken by storm by the deep learning wave since the last few years. We saw unprecedented advances in many natural language understanding tasks [21] such as text classification or sequence tagging with the rise of deeper neural

architectures, like ELMo [17] introduced in 2018 by the Allen Institute for AI.

However, one natural language understanding task still fell short of community expectations : natural language generation. Indeed, generating meaningful, coherent, and plausible textual content is quite a challenge. Only until recently, we were able to generate phrases and sentences pretty easily [8]. Paragraphs generation was only possible with some sophisticated clever tricks [11]. It is important to note that detecting falsified textual generation is much easier for a human than analyzing images. With visual content, our eyes are drowned with information. We have to analyze thousands of pixels simultaneously in order to detect suspicious artifacts. On the contrary, when we read, we rapidly analyze each character, words sequenced to form a sentence, paragraph or even a whole document. We can easily spot which word is malformed, which sentence is incoherent with the rest of the body, and so on. This is why textual DeepFakes have seen the day a little later ; models just were not good enough at generating fake news as a whole.

Recently in June 2018, OpenAI introduced their Generative Pre-Trained (GPT) architecture specifically designed to model and generate natural language [18]. It is the first large neural network able to generate fluent natural language. Then follows TransformerXL [2], which introduces a mechanism able to generate coherent paragraphs, even whole documents. Finally, in the hunt for the best generative language model, OpenAI released in February 2019 GPT-2 [19], a larger version of their previous model, GPT. The interesting thing that Radford et al. introduced (but did not release it to the community at first !) is a brand new curated textual dataset of 40 GB. Their larger version of the model, once trained on that new English corpus, was so good at generating documents that they decided to publish only the scientific paper and some examples. No dataset, no code, only a pre-trained copy of the smaller version of their proposed architecture.

One example drawn from the sample is a dialogue including the former President John F. Kennedy and the journalist Amy Goodman. The topic is about the U.S. immigration and it is only by looking carefully at the generated text that we can spot suspicious artifacts like repetitions, a common problem inherent in language models. An important thing that we cannot identify strictly with the generation is that John F. Kennedy was assassinated on November 22, 1963 [8] and Amy Goodman was born on April 13, 1967 [9]. Amy Goodman would have been four years old during

this interview. This means that without the global context, somebody who does not know John F. Kennedy or Amy Goodman could have thought that this interview really happened.

One may ask himself, what is the usefulness of a model that is able to generate textual content ? Well, it can be a powerful tool to writing aid, as the HuggingFace company demonstrated with its online editor [10]. This tool is designed to suggest the next sequence of words given a previous input entered by the user. While it is not directly intended for automatic fake news generation, it can be a great accelerator for the person that is writing it.

A natural language generator becomes even handier if we can especially control the linguistic style aspects. Ficler and Goldberg showed that a neural architecture can be designed specifically for generating content by controlling the theme (plot, acting, production, effects), the sentiment (positive, negative or neutral) or the length (short, medium, long) of a generated piece of text. Recently, Salesforce proposed a model [10] inspired by GPT-2 that is conditioned on a variety of control elements that make desired features of generated text more explicit where one can specify the domain, style, topics, dates, entities and relationships between entities, for example. This capacity of "controllable text generation" gives one the possibility to rapidly generate a document (possibly fake news !) according to an underlying message, theme or idea that this person wants to convey which justifies the worrying aspect of these recent advances in automatic text generation.

Similar to the case of visual content generation, researchers have studied the issue of detecting fake textual news generated by artificial intelligence. Zellers et al. demonstrated that a model with lower capacity and trained on a shorter period of time than a deep fake neural language generator is able to identify up to 80% of the time a real from fake news. This means that their model will miss 20% of the fake news that lie on the internet, which can be a lot ! The assumption behind this experiment is that a detector will always be behind the latest research conducted in natural language generation, hence the difference in capacity between the detector and the generator. Nguyen et al. framed fake news detection as an inference problem in a Markov random field by leveraging correlations among news articles rather than only consider each news article individually. They applied their method on shorter utterances of text like Twitter or the Weibo dataset [13] which proved to be effective but still, improvements have to be considered in order to fight the automatic textual fake news generation.

CONCLUSION AND FUTURE DIRECTIONS

We saw in the previous section that machine learning models can already be used to generate fake content. However, these models need a lot of good data in order to obtain performances that will fool the human eye. To date, they require qualified people and are costly to train. This cost is subject to reduce with time, leaving these powerful tools within more people's reach.

We are not yet at the stage of fully automatic fake content generation, but it is getting easier to generate such falsified numeric content. This is why it is important to be aware of the capacity of machine learning models and where artificial intelligence stands in this issue that is taking the social media by storm. The number of deep fakes on the internet is intractable but it is a growing concern that will have inevitably a profound impact on multiple facets of our society.

Even though there is some work done, in order to automatically identify falsified content with machine learning models, the end consumer has the responsibility to verify a content's sources. It is important as a community to flag and report potential fake news, and to share reliable, trustworthy content.

While uncovered in this chapter, voice spoofing [14] constitutes an interesting target to artificial intelligence pirates. In September 2019, a CEO was scammed out of $243,000 by a voice deep fake model. It is, according to this article 11, "the first noted instance of an artificial intelligence generated voice used in a scam."

REFERENCES

[1] Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), 2019, pp. 1205-1207, doi : 10.1109/ICCVW.2019.00152.

[2] Dai, Zihang, Yang, Zhilin, Yang, Yiming, Carbonell, Jaime G., Le, Quoc Viet, Salakhutdinov, Ruslan, « Transformer-XL : Attentive Language Models beyond a Fixed-Length Context », *ACL* (1) 2019 : 2978-2988

[3] Ficler, Jessica and Goldberg, Yoav, "Controlling Linguistic Style Aspects in Neural Language Generation," *CoRR* abs/1707.02633 (2017)

[4] Goodfellow, Ian G, Bengio, Yoshua, and Courville, Aaron C, "Deep learning," *Nature*, 521 :436–444, 2015.

[5] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua, "Generative adversarial nets." NIPS, 2014, 2672-2680.

[6] Guera, David and Delp, Edward J., "Deepfake video detection using recurrent neural networks." 2018 15th *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pages 1–6.

[7] Hochreiter, Sepp, and Schmidhuber, Jürgen, « Long short-term memory », *Neural Computation*, 9, 1997, 1735–1780.

[8] Jurafsky, Dan and Martin, James H., *N-gram language models. In Speech and Language Processing*, Third Edition, Prentice Hall, 2018.

[9] Karras, Tero, Aila, Timo, Laine, Samuli and Lehtinen, Jaakko, "Progressive growing of gans for improved quality, stability, and variation," *ICLR*, 2018, https ://arxiv.org/abs/1710.10196v1.

[10] Keskar, Nitish Shirish, McCann, Bryan, Varshney, Lav R., Xiong, Caiming and Socher, Richard, "Ctrl : A conditional transformer language model for controllable generation," *ArXiv,* 2019. https ://arxiv.org/abs/1710.10196.

[11] Krause, Jonathan, Johnson, Johanna E., Krishna, Ranjay and Fei-Fei, Li, "A Hierarchical Approach for Generating Descriptive Image Paragraphs," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 3337-3345, doi : 10.1109/CVPR.2017.356.

[12] Li, Yuezun, Yang, Xibei, Sun, Pu, Qi, Honggang and Lyu, Siwei, "Celeb-df : A new dataset for deepfake forensics," *ArXiv*, 2019. https ://arxiv.org/abs/1909.12962

[13] Ma, Jing, Gao, Wei, Mitra, Prasenjit, Kwon, Sejeong, Jansen, Bernard J., Wong, Kam-Fai, and Cha, Meeyoung, "*Detecting rumors from microblogs with recurrent neural networks*," *IJCAI,* 2016. https ://www.ijcai.org/Proceedings/16/Papers/537.pdf

[14] Mukhopadhyay, Dibya, Shirvanian, Maliheh, and Saxena, Nitesh, "All your voices are belong to us : Stealing voices to fool humans and machines," *ESORICS*, 2015. https ://link.springer.com/chapter/10.1007/978-3-319-24177-7_30.

[15] Nguyen, Duc Minh, Do, Tien Huu, Calderbank, A. Robert, and Deligiannis, Nikos, "Fake news detection using deep markov random fields," *Proceedings of NAACL-HLT* 2019, pages 1391–1400. https ://www.aclweb.org/anthology/N19-1141.pdf.

[16] Nguyen, Thanh Thi, Nguyen, C. M., Nguyen, Dung T., Nguyen, Duc Thanh, and Nahavandi, Saeid, "Deep learning for deepfakes creation and detection," *ArXiv*, 2019. https ://arxiv.org/pdf/1909.11573.pdf.

[17] Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke, "Deep contextualized word representations," *ArXiv*, 2018. https ://arxiv.org/abs/1802.05365.

[18] Radford, Alec, "Improving language understanding by generative pre-training," *Computer Science*, 2018. https ://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[19] Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, and Sutskever, Ilya, "Language models are unsupervised multitask learners" *Computer Science*, 2019. https ://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[20] Suwajanakorn, Supasorn, Seitz, Steven M., and Kemelmacher-Shlizerman, Ira, "Synthesizing Obama : learning lip sync from audio," *ACM Transactions on Graphics*, Article No. 95, July 2017. https ://doi.org/10.1145/3072959.3073640

[21] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. "Glue : A multitask benchmark and analysis platform for natural language understanding," *ICLR* 2019. https ://arxiv.org/abs/1804.07461

[22] Zellers, Rowan, Holtzman, Ari, Rashkin, Hannah, Bisk, Yonatan, Farhadi, Ali, Roesner, Franziska, and Choi, Yejin, "Defending against neural fake news," *NeurIPS* 2019, ArXiv. https ://arxiv.org/abs/1905.12616.

[1] https ://www.merriam-webster.com/words-at-play/the-real-story-of-fake-news

[2] Elle Hunt wrote a great article about the different types of Fake News as well as the motivations behind them.

https ://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

[3] We refer the reader to the book of Deep Learning [4] for more details on the topic.

[4] Training on 8 Tesla V100 for 4 days on Amazon Web Services costs about $700 US using Amazon Spot Requests.

[5] https ://www.youtube.com/watch ?v=cQ54GDm1eL0

[6] https ://ai.facebook.com/blog/deepfake-detection-challenge/

[7] https ://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[8] https ://fr.wikipedia.org/wiki/John_Fitzgerald_Kennedy

[9] https ://fr.wikipedia.org/wiki/Amy_Goodman

[10] https ://transformer.huggingface.co/

8

PROTECTING ONLINE COMMUNITIES FROM HARMFUL BEHAVIORS

Marc-André Larochelle, Éloi Brassard-Gourdeau, Zeineb Trabelsi, Richard Khoury, Sehl Mellouli, Liza Wood, Chris Priebe

Marc-André Larochelle has obtained his bachelor's degree in computer science in 2017. He is now pursuing his master's degree at Université Laval (Québec, QC) on cyberbullying detection with the support of Two Hat Security. His research interests are deep learning models, natural language processing and big data analysis.

Éloi Brassard-Gourdeau is a data scientist at Two Hat Security since 2019. He received his Bachelor's Degree in Mathematics and Computer Science in 2017 and his Master's Degree in Computer Science, with a focus on natural language processing for harm detection, in 2019, both at Laval University (Québec City, QC). He has worked on automatic harm detection for over two years both from an academic and a business standpoint.

Zeineb Trabelsi is a third-year PhD student in Information System Department at Laval University and an Intern in the Natural language processing department at Two Hat Security. Her doctoral research interests include online game communities and focus more especially on the classification of deviant behaviors. She takes a multidisciplinary approach that encompasses sociology, information system and computer science to develop a better understanding to classify the online deviant behaviors.

Richard Khoury received his Bachelor's Degree and his Master's Degree in Electrical and Computer Engineering from Laval University (Québec City, QC) in 2002 and 2004 respectively, and his Doctorate in Electrical and Computer Engineering from the University of Waterloo (Waterloo, ON) in 2007. From 2008 to 2016, he worked as a faculty member in the Department of Software Engineering at Lakehead University. In 2016, he moved to Université Laval as an associate professor. Dr. Khoury's primary areas of research are data mining and natural language processing, and additional interests include knowledge management, machine learning, and artificial intelligence.

Sehl Mellouli is a professor at the information systems department of Université Laval (Québec, Canada). Research interests of Professor Mellouli are mainly related to smart cities, smart government, citizen participation, textual data analysis, block chains, and artificial intelligence in organizations. He has publications in highly ranked and well-known journals and conferences. Professor Mellouli holds a Ph. D in computer science from Université Laval.

Liza Wood is Director of Research and Data Science at Two Hat Security, where her team is working on the latest state of the art in the emerging field of online harm detection. She has over 13 years of experience in video game development, the last two of which were as Executive Producer of Disney's Club Penguin Island. Prior to that, she was in various engineering and project management roles in the telecommunications industry. She has a Master of Data Science from the University of British Columbia.

Chris Priebe founded Two Hat Security in 2012 and serves as their CEO. Two Hat processes over 30 billion human interactions (chat, usernames, photos) every month for many of the largest games and social sites. He was one of the keynotes for the Fair Play Alliance GDC sessions and has spoken at many other conferences on harm and safety online. He has over 20 years of experience in building online communities including Disney's Club Penguin with over 300 million users.

Online communities abound today. Some of these communities are "healthy" and foster polite discussion between respectful members, but others are "harmful" and devolve into cyberbullying, hatred, and worse. In this chapter, we will explore the topic of online harm and the challenges it poses. We will start by defining the problem of online harm, which we will show is a challenge in and of itself due to the wide variety of behaviors that share that label. We will then discuss the datasets that are publicly available to train online harm detection systems. Next, we will present two variations of the challenge of harm detection, namely single-line harm detection to determine if a single message contains harmful content, and preemptive detection from conversations to predict if an upcoming message contains harmful content. The chapter will then conclude with a discussion of community protection strategies.

PROBLEM AND CONTEXT

Online communities abound today, arising on social networking sites, on the websites of real-world communities like schools or clubs, on web discussion forums, on the discussion boards of video games, and even on

the comment pages of news sites and blogs. These communities are fundamentally social, allowing their participants to connect and communicate with like-minded individuals and build social capital (McMahon 2015). People long to share, to find validation, to find places of belonging. Each human interaction, even as small as hello or LOL or :), is a chance for people to connect.

Some of these communities are "healthy" and foster polite discussion between respectful members, but others are "harmful" and devolve into virulent fights, trolling, cyberbullying, fraud, or worse even, incitation to suicide, radicalization, or the sexual predation and grooming of minors. Harmful comments are unfortunately quite widespread today. For instance, in Canada, 17% of internet users experience cyberbullying, and the problem affects disproportionately women (19%), low-income people (24%), and homosexuals (34%) (Hango 2016). These statistics increase dramatically when it comes to online gaming : 74% of gamers experience some form of cyberbullying, and in 53% of cases they are targeted because of their race, religion, gender, or sexual orientation (ADL 2019). The social toll of online harm cannot be overstated : 10% of people develop depressive or suicidal thoughts as a result of being exposed to harmful content(ADL 2019).

Many communities rely on community guidelines and human moderation to protect community members from online harm. By displaying the rules of acceptable conduct within a community, the number of newcomers complying to those community norms increases by more than 8% and their participation rate in discussions increases by an average of 70% (Matias 2019). In addition, making community norms visible reduces harmful conversations by influencing how people behaved within the conversation and by influencing newcomers' decision to join (Matias 2019). The most common method for dealing with harmful messages is by having human moderators monitor the community and remove the problematic messages. However, even a moderately sized community will generate thousands of new messages per day, while a large social network has billions of new posts daily, which makes it impossible for moderators to read every message.

Consequently, this approach relies on community members reporting messages to the moderators for action. This makes it a very slow process, as harmful messages must be read, reported, and reach the top of the moderation queue before they are removed. In the meantime, they negatively affect the community and may even be shared and distributed

more widely. Moreover, human moderators are necessarily exposed to a steady stream of harmful messages they need to judge, which has demonstrated negative impacts on their health, including causing secondary traumatic stress and PTSD (Newton 2019).

This makes automated filters very interesting. Algorithms can easily keep up with the high rate of messages being posted and immediately block detected harmful messages, and they remove the need to expose humans, readers or moderators, to the harmful content. However, most of such software are variations of keyword detection systems, and the keyword list they rely on can quickly become impossibly long and complex. For example, to prevent users from discussing male genitalia, it is not enough to filter out the word "penis," but one must also include all popular slang nicknames ("cock," "dick," etc.). Then, one must include all alternative spellings of the words, both accidental and deliberate ("peniss," "c0ck," etc.). And then one must include all synonymous multi-word expressions ("trouser snake," "one-eyed monster," etc.), each of which is composed of acceptable individual words, along with all their spelling variants. And the list must be constantly updated, to take into account all the new ways users will come up with to circumvent this filter. Moreover, such software is not very sophisticated, and can accidentally stifle legitimate discussions, such as on male sexual health in our example.

DEFINITIONS

The first challenge when dealing with online harm is defining the problem. In fact, the definitions of what constitutes harmful messages online are very varied both in practice and in the scientific literature and can differ from person to person based on age, culture, experiences, and personal beliefs. In the absence of a clear definition, moderation decisions can appear unfair and arbitrary to users, and consequently illegitimate.

While the expression "online toxicity" is a popular label for this phenomenon online and in the scientific literature, in recent years this label has been attached to a widening array of behaviors and as a result has become vague, overloaded, and ill-defined. As a result, industry has gradually been adopting the expressions "online harm" and "disruptive behavior," which better describes the common point of all these problematic behaviors, since they disrupt the community and harm targeted individuals. Disruptive behavior also naturally covers behaviors other than messaging,

such as attacking one's teammates in an online game or doxing someone. Meanwhile, in the legal and political world, the UK government has been pioneering the use of the expression "online harm" to describe disruptive behavior along with criminal behavior in their recent work to legislate a duty of care for online community hosts (UK 2019).

Perhaps the most common and well-known form of online harm or disruptive behavior is cyber aggression and cyberbullying. The formal definition of cyberbullying is an aggressive and intentional conversation act carried out repeatedly over time against a victim who cannot easily defend themselves (Best et al. 2014) (Watts et al. 2017) (Hinduja et al. 2010), (Kowalski et al. 2014), while cyber aggression is the same act done only once (Hosseinmardi et al. 2016), (Chatzakou et al. 2017). But this definition is broad, and some authors have narrowed their definition of cyberbullying to more precise sets of behaviors, such as ridicule, slander or insinuations (Ptaszynski et al. 2015), flooding, masquerade, flaming, trolling, harassment, threats, denigration, outing or exclusions (Bayzick et al. 2011), hate speech and racism (Chatzakou et al. 2017) or aggressive and hurtful messages (Van Hee et al. 2018). Moreover, the distinction between cyber aggression and cyberbullying, namely repetition of the behavior, is often difficult to do in practice, and impossible to do when considering a single comment outside of the context of the conversation in which it occurs. As a result, many definitions confuse the two (Van Hee et al. 2018) (Ptaszynski et al. 2015), and one paper showed that most acts labelled as cyber aggression were also labelled as cyberbullying in a real-world dataset (Hosseinmardi et al. 2016).

A related category of harm is online hate speech. While this category is sometimes included with cyberbullying, it is in fact a different type of behavior, defined as "a hostile and malicious speech which has a biased motive and is directed towards a group of people because of some innate characteristics" (Alam et al. 2016). Online hate thus includes behaviors such as racism, homophobia, misogyny, anti-Semitism, and more. It is worth nothing that the boundary between hate speech and free speech, extremism, and terrorism are unclear (Alam et al. 2016) and frequently hotly debated, and as a result many differences in definitions and labelling have arisen (Salminen et al. 2018).

A form of online harm is related to online sexual activities (OSA). This broad category of behaviors can be subdivided into two subcategories. The first is sexting, which is a transmission (sending, receiving, forwarding,

asking for, etc.) of sexual or erotic content (Gámez-Guadix and Mateos-Pérez 2019) (Gámez-Guadix et al. 2015) (Acar, 2016) between adults who may or may not be consenting. The second subcategory is online child sexual victimization, which is a non-consensual OSA between adults and minors. This includes both online grooming, in which the adult tries to gain trust of minors to ensure their compliance in sexual activities (Gámez-Guadix and Mateos-Pérez 2019) (Whittle et al. 2013) (Machimbarrena et al. 2018), and sextortion, where the adult use extortion or threats against another adult or a minor to force their participation (Acar 2016).

Self-directed harm, or messages about harm to oneself, constitute the last category. This category includes messages describing non-suicidal self-harm behaviors such as cutting, burning, scratching, non-fatal poisoning, and suicide ideation (Marchant et al. 2017) (Patchin and Hinduja 2017) (Rasmussen and Hawton 2014) (Emma Hilton 2017) (Zdanow and Wright 2012) as well as suicide-related messages (Marchant et al. 2017), (Cavazos-Rehg et al. 2016), (Biddle et al. 2016). Even though the author is also the target of these messages, their negative impact on the community should not be minimized, as studies have shown a strong link between online exposure to self-harm behaviors and engaging in these behaviors (Best et al. 2014).

For the sake of simplicity and consistency, we will continue using the expression "online harm" in this chapter.

ONLINE HARM DATASETS

A key challenge in online harm research area is the availability of suitable datasets, which are necessary in order to study harmful behaviors and to train systems to automatically discover these behaviors. Indeed, many communities refuse to share their data for various reasons, such as to protect their users' privacy, because the community is volatile and the data is never stored, or because the company considers the data valuable and wants to exploit it for profit. Regulations such as Europe's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act require online platforms to only use the data for purposes for which the platform user gives permission. In addition, the users can request their data be deleted the data, changing the baseline and making it very difficult automatically train systems with the data (Greene et al. 2019). With fines of €20 million for violations against these regulations, companies are

hesitant to share even for the greater good. As a result, only a small number of communities make their data publicly available for research. To make matters worse, harmful behaviors are the exception in communities, and as a result these datasets are very imbalanced against harmful behaviors, sometimes featuring only 5% of harmful messages (Emmery et al. 2019), unless such messages has been artificially over-sampled. This means that harmful messages are actually quite rare, even in online harm datasets.

Identifying and labelling online harm behaviors in the datasets is also a challenge. There is no dataset of all types of harmful behaviors we listed in the previous section; each dataset specializes on a specific subset of it. But even within a given subset, there can be a lot of ambiguity and subjectivity when defining what distinguishes certain harmful behaviors from others (Founta et al. 2018). As a result, a dataset labelling can end up reflecting its creators' biases in what they consider harmful. To further add to the challenge, most communities allow several different types of interactions simultaneously, such as public and private messages, voice chat, image and video sharing, and gaming, but datasets do not include this full range of activities and will thus necessarily give an incomplete view of the social interactions going on.

The authors of (Emmery et al. 2019) and (Larochelle and Khoury 2020) have noted that there is minimal vocabulary overlap between the various publicly available online harm datasets. This has led them to call into question the ability of harm detection models to generalize to corpora they weren't trained on. To demonstrate the issue, (Emmery et al. 2019) trained a SVM on one online harm message corpus and showed that its F1-score dropped by 15% to 30% when tested on a different corpus. Likewise, (Larochelle and Khoury 2020) trained a deep neural architecture using one of eight different online harm datasets and tested it on the other seven datasets, and showed that precision and recall fluctuated greatly and unpredictably from one test to the next. As a result, a system trained with one corpus can be of limited usefulness for detecting harmful messages in a real environment.

Nonetheless, there are several datasets publicly available and used in research. For single messages, some of the most popular datasets include a Twitter dataset (Chatzakou et al. 2017) of 9,484 individual tweets labelled to identify aggressive and abusive messages, and a dataset of 115,737 messages from Wikipedia's talk pages (Wulczyn 2017) labelled for personal attacks. There are also two datasets that were also created by Jigsaw for

Kaggle competitions, one composed of 160,000 Wikipedia messages (Kaggle 2017) and the other of 1.8 million messages from the now-defunct Civil Comments platform (Kaggle 2019). Both are labelled with six categories of cyber aggression and hate speech, and the second is considered one of the largest, if not the largest, online harm message datasets available.

In fact, most datasets available today are of cyber aggression and hate speech messages. Few datasets exist for the other types of online harm. One that is worth mentioning is the dataset of 262,673 PornHub comments taken from the most viewed videos on the site (kayleighhappersett 2019). This is a resource rich in OSA-type harm, and the comments are made in a variety of languages (by contrast, other toxicity datasets are in a single language, usually English). A quick scan of the dataset reveals that 74% of the comments are in English, and of those 46% contain obvious harmful language.

There also exists a few datasets of entire conversations. For cyberbullying, there is a dataset of 139 MySpace conversations (Ventirozos et al. 2017), each one composed of between 7 and 48 messages, with portions of each conversation marked as containing cyberbullying. The Perverted Justice website has also made available a dataset of chat conversations (Kontostathis 2009) between volunteers posing as teenagers and sexual predators. There are 259 such conversations, ranging between 31 and 22,207 messages, and while these harmful conversations were all reported to the police and led to arrests, the messages they are composed of are not labelled for toxicity.

SINGLE-LINE HARM DETECTION

The most common form of harmful message detection is single-line detection. In other words, each message is rated for risk of harm by itself, outside of the context of the surrounding messages.

As explained earlier, one limitation of keyword-based filters is their inability to handle ambiguous words that may be harmful in some context but healthy in others. The authors of (Ando et al. 2010) tried to resolve this issue by taking into account the context in which these words occur. They start by defining a list of ambiguous words, which they call gray words. When one of these gray words is detected in a message, they consider alongside the other words of the message and compute the probability of these words co-occurring in harmful or healthy messages. Their results

show that two-word co-occurrences correctly disambiguate a gray word as harmful or not with 50% precision, and three-word co-occurrences does so with over 90% precision.

Words embeddings are a popular representation of language. They are trained to convert the high-dimensional vector space of language into a dense representation in which words with similar meaning will cluster close together and dissimilar ones will be far apart, providing dense representation of each word. In (Rakib and Soon 2018), the authors decided to train a domain-specific word embedding using a corpus of Reddit posts. They found that using this new word embedding improved the precision of cyber aggression detection by 2% compared to using non-domain-specific word embedding, and by up to 12% compared to using traditional words-based features directly.

In light of the issues with the generalization of single-line detection tools described in (Emmery et al. 2019) and (Larochelle and Khoury 2020) and mentioned in the previous section, a shift has been made recently towards using language understanding models. Their intrinsic language representation enables them to perform better in this task. In fact, a recent study on the detection of offensive language in social media (Zampieri et al. 2019) found that the three best systems for the task were all based on deep learning transformers, such as BERT (Devlin et al. 2019) and GPT (Radford et al. 2018), which could also help mitigate the problematic biases induced by the data, as shown by (Borkan et al. 2019).

As this section shows, single-line harm detection has come a long way, from systems that tried to improve on keyword detection a decade ago to deep-learning language understanding systems proposed today. These latest developments are paving the way for an improved detection that will be able to catch harmful messages that do not contain explicitly profane, vulgar or obscene language. Such systems can serve a vital role in an online community, immediately filtering out clearly harmful messages and tagging potentially unsafe comments for review, and thus helping community moderators create a safe and positive environment for users. However, single-line detection will always be limited due to the lack of conversation context (Emmery et al. 2019). Indeed, a positive message can be used as provocation in a debate, and an insult message can be banter between friends. Any system that considers those messages outside of their context will lack the crucial details needed to correctly interpret them. This will be the focus of the next section.

CONVERSATION HARM PREDICTION

Although most work on harm detection has been focused on single message content, many forms on online harm arise from conversation, including notably cyberbullying and online child sexual victimization. These kinds of behavior cannot be detected by looking at individual single-line messages and require the analysis of an entire conversation to detect. In addition, handling an entire conversation instead of individual messages makes it possible to do more than simply detect harmful messages. By monitoring how a conversation is evolving, it becomes possible to predict whether harmful messages will be posted in the future. This task is known as preemptive detection. Its benefit to online moderation should be clear : such a system would allow moderators to intervene in a conversation before it degrades out of control and to rein in users before they start attacking each other, thus creating a much more positive online experience both for these users and for the other participants in the community.

One of the first studies on preemptive detection is that of (Zhang et al. 2018). In their work, the authors try to find features in the first two messages of a conversation that indicate a higher probability of it going awry. The conversations in this study, are very short and are taken from a variety of Wikipedia talk pages. The conversation features analyzed are both politeness strategies and rhetorical prompts. There are two types of politeness strategies, namely positive strategies such as greetings or saying "please" and "thank you," and negative strategies such as direct questions and sentence-initial second-person pronouns. Rhetorical prompts are separated into six categories : factual checks, moderation, coordination, casual remarks action statements, and opinions. The authors of (Zhang et al. 2018) used these features to train a logistic regression model which could predict if conversations would degrade into harmful behaviors with 61.6% accuracy ; by comparison, a human reader can do this with 72% accuracy. Their regression model showed that direct questions, second person starts, and factual checks were the strongest predictors of conversations becoming harmful, while greetings and the presence of hedges were strong predictors of conversations remaining healthy.

The work of (Zhang et al. 2018) inspired the authors of (Karan, and Šnajder 2019), who ran experiments using a similar, but much bigger, dataset. They studied two different tasks, the first being pure preemptive detection as in (Zhang et al. 2018) and the second one being post hoc single-line detection using the whole conversation for context. For both

tasks, they experimented with a SVM trained on TF-IDF weighted unigrams and bigrams, and with a BiLSTM with GLoVE embedding, an architecture which has proven itself in single-line harm detection. The neural network performed better in both tasks, achieving a 62% accuracy in the preemptive moderation task and 90.4% accuracy in the post hoc detection task. However, for the post hoc task, that accuracy was only marginally better than what the system achieved in single-line detection without looking at the rest of the conversation. It is worth highlighting though that both systems used words as features, and not the most sophisticated conversation-based features of (Zhang et al. 2018), which may explain its weaker results.

The authors of (Liu et al. 2018) studied preemptive detection in the comments of Instagram posts. They extracted a variety of features from these comments, from more traditional ones such as n-grams, word embedding, and lexicons, to platform-specific ones such as past user activity and behavioral trends in the comment threads. The logistic regression model they trained achieved an F1-score of 76.5% in this task. Based on their analysis, the four most important features for preemptive detection are: (1) whether the author received hostile comments in the past; (2) the presence of user-directed profanity; (3) the number of different users participating; and (4) trends in hostility so far in the conversation.

Finally, the authors of (Brassard-Gourdeau and Khoury 2020) studied the impact of sentiment information for preemptive detection. They did this first by reproducing the work of (Zhang et al. 2018) and adding sentiment information as a feature to their regression model, and second by training a neural network like the one of (Karan, and Šnajder 2019) to do preemptive detection on a new gaming chat conversation dataset. Their results show that sentiment information is indeed a predictor of upcoming health or harm, but only when considered at word-level granularity; trying to compute a message's average sentiment instead wipes out that benefit. Moreover, their study showed that a lot of positive words are needed to keep a conversation healthy, but only a couple of negative words can turn a conversation harmful very quickly.

To summarize, while the task of preemptive detection is a new challenge, we can already see some guiding principles start to emerge from the studies done so far. In all the studies so far, the best results have come from looking at user interactions, both directly in the case of (Liu et al. 2018) and indirectly through their language in the case of (Zhang et al.

2018) and (Brassard-Gourdeau and Khoury 2020), and the most predictive features have been very fine-grained, such as the use of specific pronouns in (Zhang et al. 2018) or the sentiment value of individual words in (Brassard-Gourdeau and Khoury 2020). Looking forward, we expect that this avenue of research will continue to develop ways of modeling human interactions from through message threads, and as the accuracy of these models improve, so will their ability to predict ahead of time which conversations are trending towards dangerous territory.

COMMUNITY PROTECTION STRATEGIES

While the machine-learning filters, we have presented in the previous sections can be very efficient at picking out harmful messages online, they do have some weaknesses. Most notably, they require a large corpus of labelled training data to learn their task. While this is true of most learning algorithms, it is particularly a problem for harm detection due to the fast-changing nature of online conversations. This means that a new insult or behavior cannot be detected by such a system, since it was never part of the training data corpus. It must be picked out and labelled manually to be integrated into the training corpus, and be observed sufficiently frequently in order to rise above the level of noise in the corpus, before the filter can be retrained to detect this new example of online harm.

In fact, a complete harm filtering strategy must necessarily be a multi-layer approach. The first layer is simply to let users know what is expected behavior in the community and to enforce these rules. This may seem obvious, but we are social beings and we adapt our behavior to what we perceive are the social norms. People behave differently at a playground, at a religious service, and at a pub, and likewise online users behave differently in communities that have clear rules and rigorous moderation compared to "free-for-all" communities (Matias 2019). When a user violates the guidelines he agreed to when he joined the community, the system should remember that infraction and build a reputation for the user. However, this reputation should not only go down to punish infractions, but also go up to reward good behaviors and create a positive incentive to respect the rules.

The next layer of defense can be keyword detection filters. While naive and simple to circumvent, these do have the advantages of being very fast to execute, and of systematically identifying the most egregious cases of

online harm and the most obviously safe messages. In the context of having to sort through millions of daily messages or more, they do provide a low-computational-cost line of defense. Keyword filters can be enhanced by decision rules to recognize common strategies to avoid the filter or avoid common false positives, at little additional overhead cost. These filters can be quickly updated by adding new keywords and rules to react instantly to the emergence of new types of online harm. These new additions can serve as a stopgap until the machine learning filters can be retrained, after which point, they should be re-evaluated to avoid the keyword and rule lists growing unwieldy.

More intelligent machine-learning analysis can then be applied as a final line of defense on messages that the filters could not whitelist or blacklist reliably. These algorithms are ideal for dealing with more complex cases, and since they are only applied to a subset of all posted messages the algorithms can use more time and resources to process the content than they could if they had to examine all messages systematically. Given the wide variety of types of online harm, multiple algorithms should be trained and specialized to reliably pick out a specific type of behavior. Messages could be triaged into one or the other of these algorithms based on what type of harmful behavior they are suspected of containing, or they could go through an ensemble model that combines all these algorithms.

CONCLUSION

In just over two decades, we have gone from having a few basic online social communities to having a rich experience. Most users today are part of multiple communities, from general ones to connect with friends and families to specialize ones based on their personal interests. Every one of these communities is a chance for a person to connect, to exchange ideas, to develop friendships. Unfortunately, it is also a risk to be exposed to bullying, hate, unwanted sexual attention, to become more disconnected, isolated and alone. Online harm is not only measured but it is felt.

In this chapter we have sought to summarize the latest state of the art in the emerging field of online harm detection. Much like the problem it seeks to correct, this research area is still new. A lot of work remains to be done, from properly defining the many facets of the issue to creating datasets that are open and representative of the real world. Technology is advancing, but online harm will not be solved by a single solution. We need

to design a multi-layer approach that blends traditional solutions with new technologies, that rewards good behaviors as well as prevents harmful ones, and helps moderators monitor their communities and users thrive in them.

Many of the online harm detection systems in the literature achieve upwards of 90% accuracy in this task, and that is great on paper. But we should not lose sight of the fact that the 10% failure cases have serious impacts on real human beings. False positives are not just statistics, they are times when friends are prevented from laughing together, a student is stopped from exploring a new idea, legitimate free speech is stifled. And every false negative is an instance where a bully was not opposed, a predator was not caught, a suicide was not prevented. Online harm is a human problem.

REFERENCES

(Acar 2016) Acar, K. V, "Sexual Extortion of Children in Cyberspace," *International Journal of Cyber Criminology*, *10* (2), 2016.

(ADL 2019) "Free to Play ? Hate, Harassment, and Positive Social Experiences in Online Games," *ADL Report*, July 2019.

(Alam et al. 2016) Alam, I., Raina, R. L. & Siddiqui, F. "Free vs hate speech on social media : the Indian perspective", *Journal of Information, Communication and Ethics in Society*, *14* (4), 350-363, 2016.

(Ando et al. 2010) Ando, S., Fujii, Y., & Ito, T. "Filtering Harmful Sentences Based on Multiple Word Co-occurrenc", *Proceedings - 9th IEEE/ACIS International Conference on Computer and Information Science*, ICIS 2010. 581-586.

(Bayzick et al. 2011) Bayzick, J., Kontostathis, A., & Edwards, L. "Detecting the presence of cyberbullying using computer software", *Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science*, 2011. http ://webpages.ursinus.edu/akontostathis/BayzickHonors.pdf.

(Best et al. 2014) Best, P., Manktelow, R., & Taylor, B. "Online communication, social media and adolescent wellbeing : A systematic narrative review," *Children and Youth Services Review*, 41, 27-36, 2014.

(Biddle et al. 2016) Biddle, L., Derges, J., Mars, B., Heron, J., Donovan, J. L., Potokar, J., Piper, M., Wyllie, C., & Gunnell, D. (2016). "Suicide and the Internet : Changes in the accessibility of suicide-related information between 2007 and 2014," *Journal of Affective Disorders*, *190*, 370-375, 2016.

(Borkan et al. 2019) Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019) "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification," *arxiv preprint*. arXiv :1903.04561.

(Brassard-Gourdeau and Khoury 2020) Brassard-Gourdeau, E., & Khoury, R. (2020) "Using Sentiment Information for Predictive Detection of Toxic Comments in Online Conversations," *arxiv preprint*. arXiv :2006.10145.

(Cavazos-Rehg et al. 2016) Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S. J., Connolly, S., Rosas, C., Bharadwaj, M., Grucza, R. & Bierut, L. J. « An analysis of depression, self-harm, and suicidal ideation content on Tumblr », *Crisis : The Journal of Crisis Intervention and Suicide Prevention, 38* (1), 44–52, 2016. https ://doi.org/10.1027/0227-5910/a000409

(Chatzakou et al. 2017) Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. "Mean birds : Detecting aggression and bullying on twitter," In *Proceedings of the 2017 ACM on web science conference* (pp. 13-22). ACM.

(Devlin et al. 2019) Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding.

(Emma Hilton 2017) Emma Hilton, C. (2017). Unveiling self-harm behavior : what can social media site Twitter tell us about self-harm ? A qualitative exploration. *Journal of clinical nursing*, *26* (11-12), 1690-1704.

(Emmery et al. 2019) Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., & Daelemans, W. (2019). Current Limitations in Cyberbullying Detection : on Evaluation Criteria, Reproducibility, and Data Scarcity.

(Founta et al. 2018) Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

(Gámez-Guadix and Mateos-Pérez 2019) Gámez-Guadix, M., & Mateos-Pérez, E. (2019). Longitudinal and reciprocal relationships between sexting, online sexual solicitations, and cyberbullying among minors. *Computers in Human Behavior*, *94*, 70-76.

(Greene et al. 2019) Greene, T., Shmueli, G., Ray, S., & Fell, J. (2019). Adjusting to the GDPR : The Impact on Data Scientists and Behavioral Researchers. *Big data*.

(Hango 2016) Hango, Darcy William, "Cyberbullying and cyberstalking among Internet users aged 15 to 29 in Canada," 2016.

(Hinduja et al. 2010) Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of suicide research*, *14* (3), 206-221.

(Hosseinmardi et al. 2016) Hosseinmardi, H., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2016). Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 186-192). IEEE.

(Kaggle 2017) Kaggle, Toxic Comment Classification Challenge. (2017) https ://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Last accessed on 10 December 2019.

(Kaggle 2019) Kaggle, Jigsaw Unintended Bias in Toxicity Classification. (2019) https ://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. Last accessed on 10 December 2019.

(Karan, and Šnajder 2019) Karan, M., & Šnajder, J. (2019). Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context, *Workshop on Abusive Language Online*, *3,* 129-134.

(kayleighhappersett 2019) kayleighhappersett Github repository. cumments. https ://github.com/kayleighhappersett/cumments. Last accessed on 10 December 2019.

(Kontostathis 2009) Kontostathis, A. (2009). Chatcoder : Toward the tracking and categorization of internet predators. In *Proceeding of the Text Mining Workshop 2009 held in conjunc-*

tion with the Ninth Siam International Conference On Data Mining (Sdm 2009). Sparks, NV. May 2009.

(Kowalski et al. 2014) Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age : A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, *140* (4), 1073.

(Larochelle and Khoury 2020) Larochelle, M.-A, & Khoury, R. (2020) "Generalisation of Cyberbullying Detection", *arxiv preprint*. arXiv :2009.01046 [cs.CL].

(Liu et al. 2018) Liu, P., Guberman, J., & Hemphill, L. and Culotta, A. (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features, *International AAAI Conference on Web and Social Media, 12*.

(Machimbarrena et al. 2018) Machimbarrena, J., Calvete, E., Fernández-González, L., Álvarez-Bardón, A., Álvarez-Fernández, L., & González-Cabrera, J. (2018). Internet risks : An overview of victimization in cyberbullying, cyber dating abuse, sexting, online grooming and problematic internet use. *International journal of environmental research and public health*, 15 (11), 2471.

(Marchant et al. 2017) Marchant, A., Hawton, K., Stewart, A., Montgomery, P., Singaravelu, V., Lloyd, K.,... & John, A. (2017). A systematic review of the relationship between internet use, self-harm and suicidal behavior in young people : the good, the bad and the unknown. *PLoS One*, *12* (8), e0181722.

(Matias 2019) Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116 (20), 9785-9789.

(McMahon 2015) McMahon, C. (2015). Why do we "like" social media ? *The Psychologist*.

(Newton 2019) Newton, Casey, "The Trauma Floor : The secret lives of Facebook moderators in America," *The Verge*, 25 February 2019.

(Patchin and Hinduja 2017) Patchin, J. W., & Hinduja, S. (2017). Digital self-harm among adolescents. *Journal of Adolescent Health*, *61* (6), 761-766.

(Ptaszynski et al. 2015) Ptaszynski, M., Masui, F., Kimura, Y., Rzepka, R., & Araki, K. (2015). Automatic Extraction of Harmful Sentence Patterns with Application in Cyberbullying Detection. In *Language and Technology Conference* (pp. 349-362). Springer, Cham.

(Radford et al. 2018) Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

(Rakib and Soon 2018) Rakib, T.B.A., & Soon, L-K (2018). Using the Reddit Corpus for Cyberbully Detection, *Intelligent Information and Database Systems*, Springer, 180-189.

(Rasmussen and Hawton 2014) Rasmussen, S., & Hawton, K. (2014). Adolescent self-harm : a school-based study in Northern Ireland. *Journal of affective disorders*, *159*, 46-52.

(Salminen et al. 2018) Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J. (2018, June). Anatomy of online hate : developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.

(UK 2019) "Online Harms White Paper," *UK Parliament report*, April 2019.

(Van Hee et al. 2018) Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13 (10), e0203794

(Ventirozos et al. 2017) Ventirozos, F. K., Varlamis, I., & Tsatsaronis, G. (2017). Detecting aggressive behavior in discussion threads using text mining. In International Conference on Computational Linguistics and Intelligent Text Processing (pp. 420-431). Springer, Cham.

(Watts et al. 2017) Watts, L. K., Wagner, J., Velasquez, B., & Behrens, P. I. (2017). Cyberbullying in higher education : A literature review. *Computers in Human Behavior*, *69*, 268-274.

(Whittle et al. 2013) Whittle, H., Hamilton-Giachritsis, C., Beech, A., & Collings, G. (2013). A review of online grooming : Characteristics and concerns. *Aggression and violent behavior*, 18 (1), 62-70.

(Wulczyn 2017) Wulczyn, E. & Thain, N. & Dixon, L. (2017). Ex Machina : Personal Attacks Seen at Scale.

(Zampieri et al. 2019) Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6 : Identifying and Categorizing Offensive Language in Social Media, *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, 75-86.

(Zdanow and Wright 2012) Zdanow, Carla, and Bianca Wright. "The representation of self injury and suicide on emo social networking groups." *African Sociological Review/Revue Africaine de Sociologie* 16.2 (2012) : 81-101.

(Zhang et al. 2018) Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., & Hua, Y. & Taraborelli, D. & Thain, N. (2018). Conversations Gone Awry : Detecting Early Signs of Conversational Failure, *Annual Meeting of the Association for Computational Linguistics, 56*, 1350-1361.

9

EXTRÉMISME VIOLENT DE DROITE ET MÉDIAS SOCIAUX : CARACTÉRISTIQUES, IDÉOLOGIES, MÉDIATISATIONS ET GESTIONS

Schallum Pierre

RÉSUMÉ

La lutte contre l'extrémisme violent de droite à travers les médias sociaux soulève des enjeux technologiques et éthiques de grande envergure. Ce chapitre montre comment les discours et les actes haineux médiatisés sur des plateformes populaires comme Facebook ou Twitter s'inscrivent dans une réappropriation de traditions idéologiques multiples d'ordre juridique, ethnique, mythique, nationaliste et identitariste. Il analyse la force et les limites des modes de gestion de l'extrémisme violent de droite qu'offrent l'intelligence artificielle et la chaîne de blocs ainsi que la pertinence d'une approche préventive impliquant les repentis.

1. INTRODUCTION

L'extrémisme violent de droite connaît une grande diffusion sur les médias sociaux. En 2017, une étude allemande révélait que 67 % des internautes ont déjà vu des commentaires haineux sur des sites Web, des blogues, des médias sociaux et des forums (Forsa, 2017 : 1). Attaquant souvent les minorités ethniques, sexuelles ou des croyances religieuses, le discours extrémiste de droite, qu'il soit haineux ou méprisant, parvient, selon Soral et alt., à créer une désensibilisation face à la violence verbale et les préjugés s'y afférant (Soral et alt., 2017). À ce titre, le discours extrémiste violent de droite soulève des enjeux rattachés à la santé communautaire, au bien-être collectif et à la sécurité publique. La diversité et la complexité de ses manifestations enjoignent les pays à réfléchir et proposer de nouvelles stratégies et de nouveaux outils en cyberdéfense. Devant la gravité de la situation, certains pays exigent que les plateformes s'engagent expressément dans la lutte contre la propagation de l'extrémisme violent sur les médias sociaux. C'est ainsi que l'« appel de Christchurch pour supprimer les contenus terroristes et extrémistes violents en ligne » (Ministry of Foreign Affairs and Trade, 2019) a été lancé par le président français Emmanuel Macron et la première ministre néo-zélandaise Jacinda Ardern, à la suite de l'attentat terroriste qui a eu lieu le 15 mars 2019 à Christchurch, en Nouvelle-Zélande. Mais, la détection grâce à l'intelligence artificielle (IA), la suppression de contenu, l'interdiction de publier, le contrôle des moindres faits et gestes des utilisateurs et utilisatrices des médias sociaux sont-ils suffisants pour endiguer le fléau de la prolifération[1] de l'extrémisme violent de droite en ligne ?

Ce chapitre traite des caractéristiques de la notion d'extrémisme violent de droite sur les médias sociaux. Il aborde son fondement idéologique et sa médiatisation, tout en mettant en lumière les grands défis technologiques et éthiques soulevés par sa gestion. Enfin, il montre la place que la prévention doit jouer dans la lutte contre l'extrémisme violent de droite.

1. L'exemple le plus récent est le cas d'extrémistes de droite allemands qui ont été arrêtés parce qu'ils « prévoyaient des attaques de grande ampleur contre des mosquées sur le modèle de Christchurch, ont révélé dimanche soir des médias allemands » (Agence France-Presse à Berlin, 2020).

2. CARACTÉRISTIQUES DE L'EXTRÉMISME VIOLENT DE DROITE SUR LES MÉDIAS SOCIAUX

L'extrémisme en ligne est le fait pour une personne de croire ou défendre, de façon radicale, une position en référence à une doctrine qui peut être d'ordre politique, culturel, religieux ou sportif. Il n'est pas forcément violent ou intolérant. L'extrémisme en ligne devient dangereux lorsqu'il franchit l'étape de la promotion de la violence (Coopération internationale et développement, 2015) ou lorsqu'il incite à la haine et au mépris de l'autre. Selon Maxime Bérubé et Aurélie Campana, l'extrémisme violent de droite se caractérise par une grande diversité d'idéologies et de structuration : ce sont principalement les révolutionnaires, les vigilants, les suprémacistes, les millénaristes, le mouvement identité chrétienne, les ultranationalistes, les survivalistes, les skinheads et les antigouvernementaux (Bérubé et Campana, 2015).

L'extrémisme violent de droite sur les médias sociaux vise à toucher à l'intégrité physique ou psychologique d'une personne ou d'une communauté via des protocoles de communication pour des raisons liées à l'ethnie, à la croyance, à l'âge, au choix sexuel ou alimentaire. Il peut être une atteinte à l'identité numérique c'est-à-dire une atteinte à la trace des données à caractère personnel qu'un individu laisse derrière lui sur les plateformes d'interactions sociales. Il s'agit d'une forme de contrôle ou de manipulation exercée sur ce que Armen Khatchatourov appelle le « soi quantifié » (Khatchatourov, 2019 : 2). L'extrémisme violent de droite sur les médias sociaux est une attaque malveillante qui peut utiliser des messages ou des images dans l'objectif de nuire à la réputation d'une personne ou d'un groupe de personnes. Ces messages ou ces images peuvent promouvoir des stéréotypes, des discours de vengeance et encourager à commettre des actes terroristes. Il se fonde sur de multiples idéologies.

3. FONDEMENT IDÉOLOGIQUE DE L'EXTRÉMISME VIOLENT DE DROITE EN LIGNE

L'extrémisme violent de droite en ligne s'inscrit dans une tradition d'interprétations et de réappropriations. Il est la résultante d'une construction de récits inspirés d'ouvrages en lien, entre autres, à des idéologies basées sur la hiérarchisation raciale, la pureté raciale et la préservation de l'identité ethnique. Il favorise l'exclusion, en référence à

des croyances racistes, mythologiques, traditionalistes, populistes et nationalistes. Cinq catégories d'idéologies seront décrites ci-après.

La première catégorie d'idéologie définissant l'extrémisme violent de droite est le racisme anti-noir tirant son origine de la période coloniale. Elle concerne la croyance de la non-appartenance des personnes noires à l'humanité. Cette conception, tirée du Code noir (Colbert, 1685), se rapporte au statut juridique d'« êtres meubles » (article 40) et de propriétés, attribué aux esclaves noirs de l'Amérique au 17e siècle. Ce statut d'être meuble des esclaves signifie également que ces derniers ne peuvent « rien avoir qui ne soit à leur maître », en référence à l'article 22 (Colbert, 1685)[2]. En cas d'excès, de vols ou de fuite, des mesures disciplinaires et dissuasives étaient prises comme la peine corporelle, le jarret coupé, les oreilles coupées, voire la peine de mort. Ainsi, s'exerce, dès cette période, un contrôle du corps noir, considéré « comme une potentielle menace à l'ordre colonial » (CNCDH, 2020 : 12). Cette conception du Noir, comme non-être à traquer et surtout comme ne faisant pas partie de l'« universalisme », n'a pas été remise en question par les philosophes des Lumières comme Rousseau ou Kant, selon Louis Sala-Molins (Calixte, Darbouze et Pierre, 2004 : 67). Les philosophes des Lumières, ayant fortement influencé la Déclaration des droits de l'homme et du citoyen de 1789 (Gaubert, 2019 : 35), la Déclaration d'indépendance des États-Unis (Martineau et Buysse, 2016) et la Révolution française (Furet, François, et Jean-Pierre Poussou, 1998), c'est ainsi qu'une définition des droits humains sans la reconnaissance d'une partie des êtres humains nous est parvenue jusqu'au 21e siècle.

La deuxième catégorie d'idéologie définissant l'extrémisme violent de droite se base sur une croyance de la supériorité raciale. Au 19e siècle, le livre *De l'inégalité des races humaines* a popularisé la notion de la hiérarchie raciale. Son auteur, Gobineau, va jusqu'à défendre que l'appellation de « barbares » associées aux « groupes inférieurs » soit un « juste mépris » (Gobineau, 1967 : 417). S'il est vrai que, dès le 19e siècle, plusieurs

2. Au-delà du legs du code noir dans le discours de l'extrémisme de droite, il serait intéressant de mettre en parallèle l'inconscient collectif venant de l'interdiction de posséder quoi que soit de l'article 22 et certaines interventions de la police en Amérique du Nord. Frantz Saintellemy, un Québécois d'origine haïtienne qui est aussi chef d'entreprise de calibre mondial, explique que : « Souvent, sa femme, Vicky, elle aussi entrepreneure dans le domaine du design et de la mode, et lui se font intercepter par des voitures de patrouille lorsqu'ils vont chercher ou reconduire les enfants à l'école ». Ainsi, « Parce qu'ils conduisent des voitures luxueuses, on n'hésite pas à les arrêter pour leur demander ce qu'ils font dans la vie pour se payer pareils carrosses » (Décarie, 2020).

intellectuels dont Anténor Firmin, avec son *De l'égalité des races humaines* (Firmin, 1885), ont rejeté la vision de Gobineau, ce dernier continue d'influencer la pensée contemporaine (Taguieff, 2008). Dans *Mon combat*, Hitler manifeste sa haine contre les « peuples étrangers » (Hitler, 19 ? : 226), comme les Tchèques, les Polonais, les Hongrois, les Ruthènes, les Serbes, Croates et surtout les Juifs (Hitler, 19 ? : 65), capables de contaminer avec le mélange, la culture allemande. L'hitlérisme valorise la protection de la race supérieure allemande et sa domination sur les races dites inférieures. Le rôle du plus fort est de dominer et de triompher, lors d'un combat. Seul le faible (Hitler, 19 ? : 149) peut ne pas être d'accord avec cette loi défendant la grandeur et la puissance de la race aryenne qu'énonce Hitler.

La troisième catégorie d'idéologie se base sur la mythologie. Elle est incarnée par la pensée d'Alfred Rosenberg et particulièrement son ouvrage *Le Mythe du XXe siècle* qui met entre autres en lumière la réapparition de la croix gammée noire, symbole de la pureté raciale, de l'esprit nordique ou encore de la « vérité organique germanique » (Rosenberg, 1986 : 641).

La quatrième catégorie renvoie au nationalisme culturel et à la publication de Benito Mussolini, *La Doctrine du fascisme*, laquelle définit le fascisme comme une conception culturelle de l'état, au centre de tout, qui joue le rôle de « transmetteur de l'esprit du peuple » (Mussolini, 1938 : 42) à travers la langue, les coutumes et la foi. Elle se rattache au respect de la tradition et de la mémoire caractérisant la nation et la patrie.

La dernière catégorie caractérise le courant identitaire ou identitariste. Elle veut endiguer le génocide blanc. L'une des figures actuelles les plus importantes est l'écrivain français Renaud Camus, auteur du livre *Le Grand remplacement*. Il soutient que les immigrés ainsi que leur religion musulmane, les « nouveaux occupants du territoire » (Camus, 2015 : 320), vont remplacer les blancs ainsi que leur religion chrétienne. Par conséquent, il « propose la "rémigration" des immigrés dans leur pays d'origine » (Camus, 2019). Dans *Whiteshift : Populism, Immigration and the Future of White Majorities*, le chercheur Eric Kaufmann attribue à la perte de l'identité blanche l'origine de l'avènement de Trump (Kaufmann, 2019). L'enjeu de la culture et de la démographie est, selon lui, plus déterminant que celui de l'économie ou de la politique. L'immigration est donc centrale pour comprendre la montée en flèche de l'extrémisme de droite.

Ces différents ouvrages publiés au fil des ans et leurs multiples interprétations constituent la trame du discours de l'extrémisme violent de droite qui circulent sur les médias sociaux. Parmi les plateformes les

plus utilisées, il y a Facebook, YouTube, Twitter, WhatsApp, Gab, Telegram et 4chan.

4. MÉDIATISATION DES TYPES D'EXTRÉMISME VIOLENT DE DROITE

Depuis la campagne présidentielle de Donald Trump, en 2016, une banalisation de la violence s'est installée tant dans les différentes sphères de la société états-unienne que sur les médias sociaux. Cette violence s'exprime à travers une défense inconditionnelle des armes, un discours haineux, une rhétorique de la division et surtout une normalisation de la violence que Samira Saramo décrit comme une métaviolence du trumpisme (Saramo, 2017). Devenu un « mouvement social », ce trumpisme est alimenté par les plateformes AlternativeRight[3] 4chan[4], Gab[5] et Breitbart[6]. Le mouvement a donné un essor considérable à la droite alternative ou alt-right, laquelle est représentée par des figures de proue comme Richard Spencer[7], Steve Bannon[8] et Milo Yiannopoulos[9]. Dans son article « Linguistic data analysis of 3 billion Reddit comments shows the alt-right is getting stronger » (Squirrell, 2017), Tim Squirrell décrit l'Alt-Right qui n'est pas une identité homogène en cinq types de troll :

- Les connards de 4chan (4chan shitposters) qui utilisent de façon extrême des insultes à connotation sexiste, raciste et antisémite. Ils utilisent des mots comme kek, Pepe, deus vult, tendies, Dieu Empereur Trump.

3. www.*AlternativeRight.com*
4. https://www.4channel.org/
5. https://gab.com/
6. https://www.breitbart.com/
7. "Spencer popularized the term 'alt-right' to describe a loosely connected fringe movement of white supremacists, neo-Nazis and other far-right extremists. On the eve of the Charlottesville rally, Spencer and others marched through the University of Virginia's campus, shouting racist and anti-Semitic slogans" (Kunzelman, 2020).
8. "Mi-août 2017, accusé de défendre une ligne très dure, défavorable aux minorités, Bannon avait dû quitter l'administration Trump après qu'une manifestation d'extrême droite eut dégénéré à Charlottesville, en Virginie, causant la mort d'une jeune femme" (Urbain, 2018).
9. « Milo is the person who propelled the alt-right movement into the mainstream,' says Heidi Beirich, who directs the Intelligence Project at the Southern Poverty Law Center, which tracks hate groups and describes the term 'alt-right' as 'a conscious rebranding by white nationalists that doesn't automatically repel the mainstream'. Beirich says she's not even sure if Yiannopoulos believes in the alt-right's tenets or just found a juvenile way to mix internet culture and extreme ideology to get attention. 'It's like he's joking: "Ha ha, let me popularize the worst ideas that ever existed" » she says. 'That's new, and that's scar' (Stein, 2016).

- Les joueurs anti-progressifs (Anti-progressive gamers) manifestent leur haine contre les guerriers de la justice sociale, les homosexuels et les féministes. Les mots les plus courants sont SJW (social justice warrior), flocon de neige, proxénétisme, tumblr, féministe, déclenchement, GamerGate et signalisation de la vertu.
- Les militants des droits de l'homme (Men's rights activists) qui sont des masculinistes défendant le droit des hommes pour la garde des enfants. On y trouve des antiféministes, des misogynes, des célibataires involontaires (incel ou involuntary celibate). Les mots-clés de cette sous-communauté sont femelle, cane, chienne, Tchad, alpha, bêta, oméga.
- Les antimondialistes (Anti-globalists) sont les partisans des théories du complot ainsi que des personnes qui les représentent comme Alex Jones, Steve Bannon, Sean Hannity. Ils se servent des mots suivants : racaille mondialiste, establishment, marionnettes, élites, maîtres, George Soros et marxiste culturel.
- Les suprémacistes blancs (White supremacists) font usage d'un langage codé et d'un racisme implicite attaquant souvent l'Islam. Les expressions suivantes sont les plus utilisées : Islam, charia, « deus vult », culture occidentale.

C'est dans la continuité de ces multiples manifestations de l'extrémisme de droite, idéologies, mouvements, trolls, qu'apparaît la figure de Brenton Tarrant, auteur des Attentats de Christchurch contre deux mosquées, à l'origine de 51 morts et 50 blessés. Avant de commettre les attaques, le terroriste a rédigé et diffusé un manifeste de 74 pages intitulé « The Great Replacement » dans lequel il se réfère à l'écrivain Renaud Camus et s'approprie le symbole du soleil noir de la SS d'Himmler (Soullier, 2019). Brenton Tarrant explique sous forme de question-réponse (Tarrant, 2019), ses croyances, son désir de mettre fin au génocide blanc et à l'immigration massive, sa haine contre les musulman.es, son islamophobie, son racisme, etc. Le terroriste a utilisé plusieurs plateformes pour médiatiser les différentes étapes de ses attaques. D'abord, il partage sur 4Chan et Twitter (Le HuffPost avec AFP, 2019) son manifeste. Ensuite, pendant 17 minutes, il diffuse en direct sur Facebook son forfait (Florent, 2019). Devenue virale, la vidéo a été supprimée par Facebook plus de 1,5 million de fois, 24 heures plus tard, après avoir été partagée sur YouTube et Twitter (R.T., 2019). L'usage des grandes plateformes pour la diffusion massive pose le problème de la gestion ou de la modération de l'extrémisme violent de droite en ligne.

5. GESTION DE L'EXTRÉMISME VIOLENT DE DROITE EN LIGNE

La gestion de l'extrémisme violent de droite en ligne est relativement récente. En effet, depuis les attentats du 11 septembre 2001, la plupart des recherches sur l'extrémisme violent se sont limitées à l'islamisme (Koehler, 2019 : 1) et au djihadisme, en accordant peu d'attention au développement de l'extrémisme violent de droite. Or, aux États-Unis, entre 2010 et 2017, des 263 actes de terrorisme intérieur, 92 étaient dus à l'extrémisme de droite et 38 au terrorisme islamiste (Koehler, 2019 : 3). Avec les médias sociaux et l'arrivée de Trump, l'extrême droite est passée d'un activisme traditionnel, hors-ligne, à un activisme en ligne (Sterkenburg, 2019 : 24) dont les impacts sont très importants. En 2016, alors que Twitter était à l'époque le média social préféré de l'état islamique, une étude comparative (Berger, 2016 : 3) révélait que les mouvements de nationalistes blancs y ont connu une augmentation de plus de 600 % depuis 2012. L'une des raisons, d'après la même étude, est que Twitter supprimait à l'époque peu les comptes des nationalistes blancs et des nazis. Tandis que les comptes des partisans de l'état islamique étaient rapidement suspendus. L'extrémisme en ligne intéresse également le gouvernement du Québec. Selon le « document relatif à » l'état de situation, des groupes d'extrême droite et de gauche actifs au Québec » du ministère de la Sécurité publique datant du 3 février 2017, les personnes qui partagent les idéaux de l'extrême droite sont actives sur les médias sociaux et se positionnent contre la montée de la religion musulmane au Québec (ministère de la Sécurité publique du Québec, 2017). L'Attentat de la grande mosquée de Québec par Alexandre Bissonnette, auquel se réfère Brenton Tarrant, confirme le document de sécurité publique. En tant que « Pro-Israël, pro-Trump, antiféministe et anti-immigrant », Bissonnette exprimait ouvertement sur Facebook (Pélouas, 2017) ses croyances. En 2017, à l'échelle canadienne, la police a déclaré une hausse de 47 % de crimes haineux comparativement à 2016 (Armstrong, 2019). Devant la montée de l'extrême droite et surtout son développement en ligne, le Canada a financé plusieurs projets de recherche afin de mieux comprendre le phénomène au pays (Sécurité publique Canada, 2020). Une charte canadienne du numérique a même été rédigée dont l'un des dix principes est le suivant : « Les Canadiens peuvent s'attendre à ce que les plateformes numériques ne servent pas à diffuser des discours haineux ou du contenu criminel, ou à promouvoir l'extrémisme violent » (Innovation, Sciences et Développement économique Canada, 2020).

Plusieurs possibilités sont envisagées dans la lutte contre la promotion de l'extrémisme violent de droite. Trois seront esquissées dans les pages suivantes. La première est la gestion par l'IA.

5.1. Gestion de l'extrémisme violent de droite en ligne par l'IA

Les exemples de Brenton Tarrant, d'Alexandre Bissonnette et de beaucoup d'autres cas d'attaque terroriste tendent à défendre l'idée que la détection de discours de haine pourrait aider à la lutte contre l'extrémisme violent de droite sur les médias sociaux. Étant donné le volume important de messages diffusés via des plateformes comme Facebook, Twitter, YouTube, il devient indispensable d'avoir des outils pouvant contribuer à l'analyse automatique de leur contenu. L'apprentissage automatique est aujourd'hui utilisé pour détecter des discours extrémistes violents qui peuvent être d'origine haineuse. Cependant, plusieurs enjeux sont à considérer, selon MacAvaney et al. comme les critères de collecte et d'étiquetage des données et l'opacité des décisions découlant d'une détection (MacAvaney and alt., 2019). Un autre défi de taille relève de la définition de l'extrémisme violent de droite. Peut-on accepter les messages qui dénotent un mépris pour une ethnie, un genre, une religion, un territoire et refuser un message qui incite à la violence liée à une ethnie, un genre, une religion, un territoire ? Où mettre la limite ? La gestion de l'extrémisme violent de droite à partir des règlements que les plateformes ont elles-mêmes rédigés pose un problème de régulation qui dépasse la sphère de la technologie. « Code is law » écrivait Lawrence Lessig (Lessig, 2000). La légifération sur les discours haineux par les états est un phénomène très récent. Elle a été laissée entre les mains des réseaux sociaux. Ces enjeux sociaux accordent aux plateformes un pouvoir insoupçonné considérable qu'il ne faut pas minimiser. Zuckerberg le reconnaît lui-même : « Par beaucoup d'aspects, Facebook ressemble plus à un gouvernement qu'à une entreprise traditionnelle. Nous disposons d'une grande communauté de personnes et à la différence de beaucoup d'entreprises technologiques, nous sommes capables d'édicter des règles » (Crawford, 2019 : 11). Si Facebook est comme une planète en soi, avec ses règles, sa cour suprême, ses juges modérateurs, c'est que la plateforme fonctionne suivant une gouvernance centralisée (Piquard, 2020). L'appel de Christchurch a montré à quel point les plateformes deviennent les équivalents des États. Ce n'était pas seulement une rencontre entre chefs d'États, mais aussi entre chefs d'États et chefs de plateformes (The Ministry of Foreign Affairs and Trade, 2019).

Cette concentration du pouvoir explique la capacité des plateformes à exercer leur influence dans différents domaines dont sur la politique. Aussi une solution qui serait à considérer est le développement de médias sociaux décentralisés.

5.2. Gestion de l'extrémisme violent de droite en ligne par la décentralisation

Un des problèmes soulevés par l'usage de l'IA pour la détection en vue de la suppression de contenu extrémiste violent, qu'il soit de droite ou de gauche, est la liberté d'expression (Rioux, 2020). De plus, la personne détectée puis bannie dont le discours vise à rabaisser, mépriser et rejeter l'autre a la possibilité de se poser en victime et d'accuser la plateforme de partisannerie. L'informatique décentralisée (Jonathan, 2020) pourrait atténuer la diffusion à grande échelle et répartir une plateforme de plus de deux milliards d'utilisateurs et d'utilisatrices comme Facebook en des millions de communautés autogérées. Il serait plus difficile pour un contenu extrémiste violent de se propager d'une communauté à une autre, sans l'autorisation de la structure d'autogestion.

Le président de Twitter est l'un des tenants de cette solution. Le 11 décembre 2019, il a annoncé la création d'une équipe (Jack, 2019), BlueSky, qui devra développer pour les médias sociaux un standard ouvert et décentralisé dont Twitter deviendrait à terme un des clients. Un protocole ouvert qui est un ensemble ouvert de règles ou de procédures régissant la communication des données entre machines (The Editors of Encyclopaedia Britannica, 2018), a l'avantage d'être transparent et rend possible la démocratisation de l'implantation. Dans *Protocols, Not Platforms : A Technological Approach to Free Speech*, Mike Masnick rappelle qu'à ses débuts Internet a été dominé par le principe des protocoles (Masnick, 2019). À titre d'exemple, les protocoles ouverts SMTP, POP3 et IMAP ont été à l'origine de multiples déploiements pour le courriel.

Un projet est déjà en expérimentation dans le domaine de la chaîne de blocs (blockchain) avec Steem (Levine, 2018) qui a créé plusieurs interfaces utilisateurs ou couches décentralisées : steemit.com, busy.org et steempeak.com. Chaque interface est régie par les règles de sa communauté, mais utilise les mêmes ressources comprenant la base de données, les utilisateurs et utilisatrices. Si un contenu extrémiste violent est interdit dans une interface, son auteur peut quand même avoir accès à son profil via les autres

interfaces. Défendant la liberté d'expression, c'est à la communauté de décider de l'avenir d'un contenu. Contrairement à Facebook qui monétise les données des utilisateurs et utilisatrices, à leur insu (Bastien L, 2019)[10], Steem récompense les personnes qui publient des contenus par un principe de vote positif ou négatif (Levine, 2018). Le contenu qui reçoit le plus de votes négatifs ne sera plus visible pour la communauté. L'utilisation de la récompense a pour objectif d'inciter les utilisateurs et utilisatrices à promouvoir une bonne conduite. Quoique le but de steem soit louable, on peut se demander l'impact que pourrait avoir une communauté dont la majorité prendrait le parti de l'extrémisme violent de droite. Les différents projets de steem, étant en développement, il est encore tôt pour déterminer l'efficacité de la gestion faite par la communauté des contenus extrémistes violents de droite. L'autogestion de l'extrémisme violent de droite exige une sensibilité à cette problématique dans le but de prévenir les actes irréversibles pour l'humanité.

5.3. Gestion de l'extrémisme violent de droite en ligne par la prévention

La gestion de l'extrémisme violent de droite par l'IA, en dépit de ses limites, est une réponse pertinente à la diffusion de contenus haineux, racistes, antisémites, homophobes et islamophobes. Elle convient tout à fait aux enjeux des plateformes comme Google, Apple Facebook et Amazon (GAFA). La gestion de l'extrémisme violent de droite par les systèmes décentralisés peut se révéler encore plus percutante. Ces stratégies peuvent être complétées par une approche préventive, ce qui exige de mettre les sciences humaines et sociales au centre du développement de la science des données. Il est tout à fait normal de se questionner sur l'autre, sur sa différence, en référence à soi. La liberté d'expression permet à tout le monde de s'interroger sur tous les sujets. Cependant, si l'on a le droit de partager son opinion sur tout, sur ce qu'on aime ou déteste, il importe de savoir que certains médiums sont peu adaptés à l'expression de ses idéaux basés sur la race, la religion, la culture et le genre. Un contenu extrémiste violent peut être un appel à l'aide[11]. En ce sens, son traitement ne devrait pas être une

10. Ce modèle d'affaires est peut-être appelé à changer avec le déploiement de la cryptomonnaie de Facebook, La Libra (Libra, 2020).
11. Les outils du Centre de prévention de la radicalisation menant à la violence (CPRMV) peuvent être très utiles : https ://info-radical.org/fr/comment-reconnaitre/

fermeture, mais une ouverture au dialogue. La détection des discours extrémistes violents ou haineux devrait être une étape devant conduire non pas à une suppression ou un bannissement, mais à un échange qui laisserait à la personne qui a des choses à raconter de s'exprimer librement. Après une détection, une stratégie visant à prendre en charge[12] les personnes formulant des contenus extrémistes violents de droite devrait être mise en place par les médias sociaux, comme un agent conversationnel dédié à la prise en charge, afin de pallier le problème. Des personnes ayant des compétences en écoute active pourraient prendre le relai. De même, les repentis comme les anciens skinheads[13], les anciens prisonniers ayant été radicalisés, pourraient, dans certains cas, jouer un rôle majeur dans la gestion de l'extrémisme violent de droite. Sur le même principe de l'embauche des chapeaux blancs (*white hat*) par des organismes publics ou privés, en vue d'améliorer la sécurité de leurs systèmes (Flood, Denihan, Keane and Mtenzi, 2012), les repentis pourraient jouer un rôle important dans la lutte contre l'extrémisme violent de droite. Ils pourraient participer à des activités de sensibilisation, de formation et d'éducation, ancrées dans le savoir expérientiel. En partageant leur expérience, les personnes ayant été radicalisées peuvent avoir des impacts considérables en mettant en lumière les différents facteurs (Centre canadien, 2018) qui les ont conduits à adopter une idéologie extrémiste : réseaux sociaux, griefs, vulnérabilités, sentiment d'appartenance, etc. La gestion de l'extrémisme violent de droite par les récits peut exiger un temps long, mais comme montre le documentaire *La Bombe* (Télé-Québec, 2018), elle peut se révéler très efficace.

6. CONCLUSION

L'extrémisme violent de droite est un phénomène complexe. Ce chapitre a voulu rendre compte d'une partie de son histoire idéologique et des outils technologiques qui concourent à sa gestion. L'IA est aujourd'hui très répandue pour la détection, laquelle peut aider à cibler de potentiels cas de violence. Il aurait été pertinent de développer aussi un enjeu éthique lié à l'exploitation des données sensibles que constitue l'expression d'une opinion fondée sur l'idéologie raciale, religieuse ou sexuelle. Ne faudrait-il pas

12. L'expérience de Moonshot pourrait être mise à contribution (Moonshot, 2020).
13. Le documentaire *La bombe* revient sur le cas de Maxime Fiset, skinhead néonazi repenti, qui « plonge dans les souvenirs refoulés de sa radicalisation d'extrême-droite, afin de mieux comprendre la tangente xénophobe que le mouvement nationaliste identitaire québécois a récemment pris » est très enrichissant (Télé-Québec, 2018).

requérir le consentement libre et éclairé d'une personne dont les données non seulement personnelles, mais aussi sensibles, soient traitées par un algorithme ? La possibilité de bannir une personne à cause de l'opinion exprimée sur un média social pose aussi le problème de la liberté d'expression. La chaîne de blocs, de son côté, s'inscrit dans la continuité du respect de la liberté d'expression. Cependant, une question qui reste ouverte est le risque que comporte le droit de vote au cas où la communauté serait contrôlée par des personnes à tendance extrémiste violente de droite. Enfin, ce chapitre préconise une gestion axée sur la prévention qui serait la garantie pour qu'une communauté soit en mesure de prendre des décisions éclairées qui vont dans le sens du respect de la personne humaine. Cependant, une réponse adéquate au phénomène de l'extrémisme violent en ligne doit tenir compte non seulement des idéologies de droite mais aussi des idéologies de gauche, du bioterrorisme et des formes religieuses et socioéconomiques. Ces questions seront traitées dans un prochain article.

BIBLIOGRAPHIE

Agence France-Presse à Berlin, « Des extrémistes arrêtés en Allemagne voulaient imiter l'attentat de Christchurch », *Le Devoir*, 17 février 2020. https://www.ledevoir.com/monde/europe/573090/des-extremistes-arretes-en-allemagne-voulaient-imiter-l-attentat-de-christchurch, consulté le 17 février 2020.

Armstrong, Amelia, « Les crimes haineux déclarés par la police au Canada, 2017 », *Statistiques Canada*, Date de diffusion : le 30 avril 2019. https://www150.statcan.gc.ca/n1/pub/85-002-x/2019001/article/00008-fra.htm, consulté le 14 juin 2020.

Bastien L, « Data Monetization : tout savoir sur la monétisation des données », *Le Big Data*, 2 avril 2019. https://www.lebigdata.fr/data-monetization-tout-savoir, consulté le 23 juillet 2020.

Berger, J. M., « Nazis vs. ISIS on Twitter :A Comparative Study of White Nationalist and ISIS Online Social Media Networks », *Program on Extremism*, September 2016, GW, p. 3. https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/Nazis%20v.%20ISIS.pdf, consulté le 14 juin 2020.

Bérubé, Maxime et Campana, Aurélie. « Les violences motivées par la haine. Idéologies et modes d'action des extrémistes de droite au Canada. » *Criminologie*, volume 48, number 1, spring 2015, p. 215–234. https://doi.org/10.7202/1029355ar

Calixte, Fritz, Darbouze, James et Pierre, Schallum. « Les Antilles : entre passé et présent. Entretien avec le professeur Louis Sala-Molins », *Recherches Haïtiano-antillaises : la figure de l'esclave noir dans le monde colonial antillais*, L'Harmattan, 2004.

Camus, Renaud, *Le grand remplacement* suivi de *Discours d'Orange*, 3e édition, Renaud Camus, collection me prévenir, 2015.

Colbert, Jean-Baptiste, *Le Code noir : Recueil d'édits, déclarations et arrêts concernant les esclaves nègres de l'Amérique*, version promulguée en mars 1685 par Louis XIV, 1685. http://www.axl.cefan.ulaval.ca/amsudant/guyanefr1685.htm, consulté le 12 juin 2020.

Commission nationale consultative des droits de l'homme (CNCDH), *Rapport 2019 sur la lutte contre le racisme, l'antisémitisme et la xénophobie : Focus sur le racisme anti-Noirs*, juin 2020. https ://www.cncdh.fr/sites/default/files/rapport_racisme_2019_focus_racisme_anti-noirs_vdef.pdf, consulté le 15 juillet 2020.

Coopération internationale et développement, *STRIVER pour le développement : renforcer la résilience face à la violence et à l'extrémisme*, Luxembourg, Office des publications de l'Union européenne, 2015. https ://rusi.org/sites/default/files/mn0115566frn.pdf, consulté le 9 juin 2020.

Crawford, Matthew, « préface » dans David, Marie et Sauviat, Cédric, *Intelligence artificielle : la nouvelle barbarie*, Monaco, éd. Du Rocher, 2019.

Décarie, Jean-Philippe, « Il faut que la règle change », La Presse, Édition du 11 juin 2020. https ://plus.lapresse.ca/screens/db96287d-d7a2-4108-a182-d80803567b85__7C___0.html ?utm_medium=Facebook&utm_campaign=Microsite+Share&utm_content=Screen, consulté le 13 juin 2020.

Firmin, Anténor, *De l'égalité des races humaines (anthropologie positive)*, Paris, Librairie cotillon, F. Pighon, successeur, imprimeur-éditeur, Libraire du Conseil d'État et de la Société de Législation comparée, 1885, 666 pp, édition numérique, pour les classiques des sciences sociales, réalisée à partir d'un fac similé de Gallica, La Bibliothèque nationale de France. Une édition réalisée par Réjeanne Toussaint, bénévole, Chomedey, Ville Laval, Québec. http ://classiques.uqac.ca/classiques/firmin_antenor/de_egalite_races_humaines/de_egalite_races_humaines.html, consulté le 2 février 2020.

Flood, J., Denihan, M. Keane, A., and Mtenzi, F., "Black hat training of white hat resources : The future of security is gaming," 2012 International Conference for Internet Technology and Secured Transactions, London, 2012, pp. 488-491.

Florent, Hugo, « Attentat. Brenton Tarrant, geek et terroriste d'extrême droite », *Courrier international,* Publié le 18 mars 2019 à 17 h 23. https ://www.courrierinternational.com/revue-de-presse/attentat-brenton-tarrant-geek-et-terroriste-dextreme-droite, consulté le 13 juin 2020.

Forsa, *Hate speech,* 2017. https ://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Service/Pressemitteilungen/Dokumente/2017/Ergebnisbericht_Hate-Speech_forsa-Mai-2017.pdf, consulté le 25 janvier 2020.

Furet, François, et Poussou, Jean-Pierre, « 8 - La philosophie des Lumières et la culture révolutionnaire », François Crouzet éd., *L'Europe dans son histoire. La vision d'Alphonse Dupront*. Presses Universitaires de France, 1998, pp. 153-167.

Gaubert, Joël, *Ernst CASSIRER, philosophe des Lumières, penseur de notre temps*, Paris, M-Éditer, 2019, p. 35.

Gobineau, Joseph-Arthur, *Essai sur l'inégalité des races humaines*, (1853-1855), Paris, éditions Pierre Belfond, 1967, 878 pages, édition numérique, pour les classiques des sciences sociales, réalisée par Marcelle Bergeron. http ://classiques.uqac.ca/classiques/gobineau/essai_inegalite_races/essai_inegalite_races.html, consulté le 2 février 2020.

Hitler, Adolf, *Mon combat,* La Bibliothèque électronique du Québec, Collection Polémique et propagande, 19 ?. https ://beq.ebooksgratuits.com/Propagande/Hitler-combat-1.pdf, consulté le 12 juin 2020.

Innovation, Sciences et Développement économique Canada, « Charte canadienne du numérique : La confiance dans un monde numérique », *Gouvernement du Canada,* Date de modi-

fication : 2020-06-08. https ://www.ic.gc.ca/eic/site/062.nsf/fra/h_00108.html, consulté le 14 juin 2020.

Jack, T*weet du 11 décembre 2019 à 9 h 13*. https ://twitter.com/jack/status/1204766078468911106, consulté le 16 juin 2020.

Jonathan, « How Decentralization Can Fight Fake News, Trolls & Other Social Media Abuses », juin 2020. https ://blog.sapien.network/how-decentralization-can-fight-fake-news-trolls-other-social-media-abuses-a4f7ba3dae11, consulté le 16 juin 2020.

Kaufmann, Eric, *Whiteshift : Populism, Immigration and the Future of White Majorities*, New York, Abrams Press, 2019. https ://www.amazon.com/Whiteshift-Populism-Immigration-Future-Majorities/dp/1468316974#reader_B07N1NCNV8, consulté le 18 juin 2020.

Khatchatourov, Armen, *Les identités numériques en tension : Entre autonomie et contrôle*, ISTE, 2019. https ://cdn.shopify.com/s/files/1/0245/3579/files/523_Les_identites_numeriques_en_tension_Katchatourov_Premiere_Partie.pdf ?4526142874484507828, consulté le 18 juin 2020.

Koehler, Daniel, "Violence and Terrorism from the Far-Right : Policy Options to Counter an Elusive Threat," *ICCT Policy Brief*, February 2019 DOI : 10.19165/2019.2.02, consulté le 14 juin 2020.

Kunzelman, Michael, "White nationalist Richard Spencer loses lawyer in lawsuit", *The Associated Press Staff Contact, CTV NEWS*, publié le 22 juin 2020 à 7 :27 PM. https ://www.ctvnews.ca/world/white-nationalist-richard-spencer-loses-lawyer-in-lawsuit-1.4995395, consulté le 15 juillet 2020.

L'Obs avec AFP, « Renaud Camus, chantre du "grand remplacement" tête de liste aux européennes », L'Obs, Publié le 09 avril 2019 à 16 h 7. https ://www.nouvelobs.com/politique/20190409.OBS11305/renaud-camus-chantre-du-grand-remplacement-tete-de-liste-aux-europeennes.html, consulté le 2 février 2020.

Le Centre canadien d'engagement communautaire et de prévention de la violence (Centre canadien), *Stratégie nationale de lutte contre la radicalisation menant à la violence*, Gouvernement du Canada, 2018. https ://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/ntnl-strtg-cntrng-rdclztn-vlnc-fr.pdf, consulté le 17 juin 2020.

Le HuffPost avec AFP, « Brenton Tarrant avait envoyé son "manifeste" à la Première ministre juste avant l'attaque », Le HuffPost, Publié le 17 mars 2019 à 5 h 36 CET et actualisé le 17 mars 2019 à 5 h 39. https ://www.huffingtonpost.fr/2019/03/17/brenton-tarrant-avait-envoye-son-manifeste-a-la-premiere-ministre-juste-avant-lattaque_a_23694048/, consulté le 13 juin 2020.

Lessig, Lawrence, « Code Is Law : On Liberty in Cyberspace », *Harvard Magazine*, 1 janvier 2000. https ://www.harvardmagazine.com/2000/01/code-is-law-html, consulté le 15 juin 2020.

Levine, Andrew, « How Steem Protects Free Speech Without Promoting Hate Speech », *Steemit*, 2 novembre 2018. https ://steemit.com/steemit/@andrarchy/how-steem-protects-free-speech-without-promoting-hate-speech, consulté le 16 juin 2020.

Libra, *Bienvenue dans le projet Libra*, La Libra Association, 2020. https ://libra.org/en-US/, consulté le 16 juin 2020.

MacAvaney, and alt., "Hate speech detection : Challenges and solutions," *PLoS ONE* 14 (8) : e0221152. Published : August 20, 2019. https ://doi.org/10.1371/journal.pone.0221152

Martineau, S. & Buysse, A. A. J. (2016). Rousseau et l'éducation : apports et tensions. *Phronesis*, *5* (2), 14–22. https ://doi.org/10.7202/1038136ar

Masnick, Mike, "Protocols, Not Platforms : A Technological Approach to Free Speech," *The Knight First Amendment Institute*, August 21, 2019. https ://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech, consulté le 16 juin 2020.

Ministère de la Sécurité publique du Québec, *Extrémisme de droite au Québec : principaux constats*, 3 février 2017. https ://www.securitepublique.gouv.qc.ca/fileadmin/Documents/ministere/diffusion/documents_transmis_acces/2017/123091.pdf, consulté le 14 juin 2020.

Ministry of Foreign Affairs and Trade, *Appel de Christchurch*, 2019. https ://www.appeldechristchurch.com/, consulté le 24 janvier 2020.

Moonshot, *We connect vulnerable individuals with Mentors*, 2020. http ://moonshotcve.com/vision/, consulté le 15 juillet 2020.

Mussolini, Benito, *La doctrine du fascisme*, Florence, 1938. https ://gallica.bnf.fr/ark :/12148/bpt6k63051t.image, consulté le 2 février 2020.

Pélouas, Anne, « Attentat dans une mosquée de Québec, l'acte d'un étudiant d'extrême droite », *Le Monde*, Publié le 31 janvier 2017 à 6 h 43 - actualisé le 31 janvier 2017 à 13 h 1. https ://www.lemonde.fr/ameriques/article/2017/01/31/attentat-dans-une-mosquee-de-quebec-l-acte-d-un-loup-solitaire_5071915_3222.html, consulté le 14 juin 2020.

Piquard, Alexandre, « Facebook réussit à réunir vingt membres pour sa "cour suprême" », *Le Monde*, Publié le 6 mai 2020 à 22 h 9. https ://www.lemonde.fr/economie/article/2020/05/06/facebook-reussit-a-reunir-vingt-membres-pour-sa-cour-supreme_6038898_3234.html, consulté le 23 juin 2020.

R.T., « Facebook a supprimé plus de 1,5 million de vidéos de l'attentat de Christchurch », *Le Parisien*, Publié e 17 mars 2019 à 20 h 56, actualisé le 20 juin 2019 à 10 h 35 https ://www.leparisien.fr/societe/facebook-a-supprime-plus-de-1-5-million-de-videos-de-l-attentat-de-christchurch-17-03-2019-8033973.php, consulté le 13 juin 2020.

Rioux, Christian, « Peut-on censurer la "haine" sur Internet ? », *Le Devoir*, 30 mai 2020. https ://www.ledevoir.com/monde/europe/579884/peut-on-censurer-la-haine-sur-internet, consulté le 15 juillet 2020.

Rosenberg, Alfred, *Le mythe du xxe siècle*, Paris, éditions Avalon, 1986. https ://archive.org/details/LeMytheDuXxeSiecle/mode/2up, consulté le 2 février 2020.

Saramo, Samira, "The Meta-violence of Trumpism," *European journal of American studies* [Online], 12-2 | 2017, document 3, Online since 10 August 2017, connection on 06 February 2020. URL : http ://journals.openedition.org/ejas/12129 ; DOI : https ://doi.org/10.4000/ejas.12129

Soral W, Bilewicz M, Winiewski M. "Exposure to hate speech increases prejudice through desensitization," *Aggress Behavior*, 2018 Mar ; 44(2) :136-146. doi : 10.1002/ab.21737. Epub 2017 Nov 2. PMID : 29094365.

Soullier, Lucie, « Attentat terroriste en Nouvelle-Zélande : ce que contient le "manifeste" rédigé par le suspect », *Le Monde*, Publié le 15 mars 2019 à 16 h 59 - Mis à jour le 16 mars 2019 à 3 h 25. https ://www.lemonde.fr/politique/article/2019/03/15/attaque-terroriste-en-nouvelle-zelande-ce-que-contient-le-manifeste-redige-par-le-suspect_5436779_823448.html, consulté le 13 juin 2020.

Squirrell, Tim, « Linguistic data analysis of 3 billion Reddit comments shows the alt-right is getting stronger », *Quartz*, August 18, 2017. https ://qz.com/1056319/what-is-the-alt-right-a-linguistic-data-analysis-of-3-billion-reddit-comments-shows-a-disparate-group-that-is-quickly-uniting/, Consulté le 13 juin 2020.

Stein, Joel, « Milo Yiannopoulos Is the Pretty, Monstrous Face of the Alt-Right », *Bloomberg*, September 15, 2016 https://www.bloomberg.com/features/2016-america-divided/milo-yiannopoulos/, consulté le 15 juillet 2020.

Sterkenburg, Nikki, *Introduction pratique à l'extrémisme de droite*, Amsterdam, Centre d'excellence du RAN, 2019, p. 24. https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/networks/radicalisation_awareness_network/ran-papers/docs/ran_fre_fact-book_20191205_fr.pdf, consulté le 14 juin 2020.

Taguieff, Pierre-André, « Figures de la pensée raciale », *Cités*, 2008/4 (n° 36), p. 173-197. DOI : 10.3917/cite.036.0173. URL : https://www.cairn.info/revue-cites-2008-4-page-173.htm, consulté le 2 février 2020.

Tarrant, Brenton, *The Great Replacemement*, 2019. https://archive.org/details/TheGreatReplacement_20190325_2009, consulté le 13 juin 2020.

Télé-Québec, *La bombe*, Blimp Télé 2 inc. 2018. https://labombe.telequebec.tv/, consulté le 17 juin 2020.

The Editors of Encyclopaedia Britannica, "Protocol," *Encyclopaedia Britannica*, August 31, 2018. https://www.britannica.com/technology/protocol-computer-science, consulté le 16 juin 2020.

Urbain, Thomas, « Descente aux enfers pour Steve Bannon, qui quitte Breitbart », *Le Nouvelliste*, 9 janvier 2018 17h18Mis à jour à 19 h 49. https://www.lenouvelliste.ca/actualites/monde/descente-aux-enfers-pour-steve-bannon-qui-quitte-breitbart-5d6e2502fd1f-78b64124420a10423b6c, consulté le 15 juillet 2020.

Ce livre dresse le portrait des médias sociaux suivant trois angles, soit la cybersécurité, la gouvernementalité algorithmique et l'intelligence artificielle. Il révèle les multiples facettes d'un sujet qui ne peut s'appréhender qu'à travers l'interdisciplinarité. Le présent ouvrage est surtout un guide pour sensibiliser le milieu citoyen aux cyberattaques et aux enjeux liés à la protection des données à caractère personnel. Il s'adresse autant aux personnes expertes en données massives qu'aux institutions privées et publiques qui rédigent des politiques de confidentialité. Nous espérons que les différents chapitres sauront contribuer à faire des médias sociaux un espace et un outil plus respectueux des données citoyennes.

SCHALLUM PIERRE est chargé scientifique et éthique à l'Institut intelligence et données (IID) de l'Université Laval et professeur à temps partiel à l'Université Saint-Paul où il enseigne le cours « Communications sociales et médias sociaux ». Chercheur en éthique des données massives, il s'intéresse à la question de l'identité dans ses dimensions technologiques, numériques, anthropologiques, idéologiques et historiques. Il a effectué un stage postdoctoral à Polytechnique Montréal dans le cadre du projet « Recherche et développement d'une plateforme de paiement mobile ». Il est détenteur d'un doctorat en philosophie de l'Université Laval et a été membre du comité d'éthique de la même université.

Dr. FEHMI JAAFAR est un chercheur au Centre de recherche en Informatique de Montréal (CRIM) et professeur adjoint affilié à l'Université Concordia. Il est le Vice chair du comité sur l'Internet des objets et technologies connexes au Conseil canadien des normes. Auparavant, il était professeur adjoint à l'Université Concordia d'Edmonton, et chercheur postdoctoral à Queen's University et à Polytechnique Montréal.Après avoir obtenu un doctorat en informatique de l'Université de Montréal, M. Jaafar s'est spécialisé dans des travaux de recherche en cybersécurité, Il s'intéresse à la cybersécurité dans l'Internet des objets et à l'application des techniques d'apprentissage automatique en cybersécurité.Il a établi des programmes de recherche en collaboration avec Défense Canada, Sécurité publique Canada, le Conseil de recherches en sciences naturelles et en génie du Canada, et des partenaires industriels et universitaires canadiens et étrangers.

Illustration de couverture : iStockphoto

ÉTHIQUE IA ET SOCIÉTÉ
Collection dirigée par Lyse Langlois

OBSERVATOIRE INTERNATIONAL SUR LES IMPACTS SOCIÉTAUX DE L'IA ET DU NUMÉRIQUE

ÉDITÉ ET IMPRIMÉ AU QUÉBEC

Communications

ISBN 978-2-7116-5328-7

Presses de l'Université Laval
pulaval.com