Transparents mais corruptibles : les algorithmes au défi des comportements « adversariaux » dans le domaine journalistique

Antonin Descampe, Université catholique de Louvain
François-Xavier Standaert, Université catholique de Louvain

RÉSUMÉ

Dans le domaine du journalisme computationnel, l'automatisation de la production de l'information invite à s'interroger sur la pertinence des décisions automatiques et les moyens de la garantir. Cette « responsabilité algorithmique » est souvent ramenée à une question de transparence, garantissant qu'un algorithme a été conçu de manière conforme à l'intention affichée. Deux études de cas impliquant une catégorisation automatique d'articles de presse montrent pourtant que la transparence n'est pas suffisante : des exemples « adversariaux » sont en mesure de détourner un algorithme de son comportement attendu. Ce constat appelle à inclure dans l'évaluation de la performance d'un algorithme la robustesse face à d'éventuels comportements « adversariaux ». Les difficultés techniques posées par cette robustesse plaident notamment pour une implication accrue des journalistes comme garants des décisions automatiques.

ABSTRACT

In the field of computational journalism, automation of information production invites us to question the relevance of automatic decisions. This "algorithmic accountability" often boils down to a transparency question, ensuring that an algorithm has been designed in accordance with its publicized intent. In this paper, two case studies involving automatic categorization of news articles show that transparency is not enough : adversarial examples are able to divert a transparent algorithm from its expected behavior. This observation suggests that robustness against adversarial behaviors should be considered in the evaluation of an algorithm's performance. The technical issues raised by this robustness argue in favor of an increased involvement of journalists as responsible for automatic decisions.

DOI: 10.31188/CaJsm.2(7).2021.R039

La disponibilité de très grandes quantités de données, combinée aux progrès récents de l'apprentissage automatique (*machine learning*), a conduit à un accroissement des décisions automatiques dans de nombreux contextes, incluant les médias d'information (Diakopoulos, 2019). Si l'automatisation peut réduire la subjectivité inhérente aux décisions humaines, elle peut également reproduire ou créer des biais et s'avérer discriminatoire. Dès lors, confier des décisions à des processus automatisés pose la question de leur capacité à rendre compte de ces mécanismes décisionnels et de l'impact de leur paramétrage. Un apprentissage automatique en mesure d'expliciter et de garantir la pertinence des décisions prises est donc devenu un sujet de recherche important, à l'intersection de l'informatique et des sciences sociales (Lepri et al., 2018). Cet apprentissage automatique « équitable » (*fair machine learning*) vise ainsi à éviter que des décisions automatiques ne reproduisent des discriminations ou des biais qui seraient déjà présents dans la société (Crawford et Schultz, 2014).

Comme le soulignent Barocas, Hardt et al. (2017), l'équité algorithmique est complexe à formaliser et donc à garantir : il n'y a pas de définition unique qui permettrait de prendre en considération l'ensemble des préoccupations sociétales soulevées par l'apprentissage automatique. Certaines caractéristiques contribuant à l'équité algorithmique sont même difficiles, voire impossibles à garantir simultanément. C'est notamment le cas de la transparence et de la robustesse, comme nous aurons l'occasion de le détailler dans la suite de ce travail. Dans le cadre large de cette équité algorithmique, nous nous focalisons dans ce travail sur le défi plus spécifique de la conception d'algorithmes responsables (*accountable algorithms*), c'est-à-dire rendant des décisions conformes à l'intention des personnes les ayant conçus et qui sont alors en mesure de se porter garants des décisions rendues.

Dans le domaine du journalisme, les enjeux techniques, organisationnels, mais aussi éthiques et épistémologiques de ce que Lewis et Westlund (2016) appellent le « *human-machine divide* » font l'objet d'une attention accrue, déplaçant le centre de gravité des *Journalism Studies* vers des objets d'étude plus proche des sciences de la technologie. À propos du journalisme computationnel, Diakopoulos (2015) formalise le défi de la responsabilité algorithmique en considérant les algorithmes comme des objets de création humaine, dont il s'agit de révéler l'intention prévalant à leur implémentation. Observant la quantité croissante de décisions que des algorithmes prennent dans les processus de production d'information, afin d'établir des priorités, de classer, associer ou filtrer des données (Diakopoulos, 2019), et du fait que « *les processus computationnels peuvent produire le contenu informationnel lui-même* » (Coddington, 2015, p. 336, traduction libre), l'enjeu de la responsabilité algorithmique revêt une importance grandissante pour bon nombre d'organes de presse qui font usage d'apprentissage automatique dans leurs processus de production.

Cette question de l'automatisation et de la responsabilité algorithmique ne se limite pas au seul journalisme computationnel et doit être considérée dans le contexte plus large de la méfiance à l'égard des médias et de la crise de confiance que de nombreuses études et baromètres de confiance rapportent (Nielsen, 2016). Ainsi, la forme particulière de pensée construite autour de l'automatisation (Wing, 2008 ; Coddington, 2015), incarnée par des interactions spécifiques entre « *acteurs sociaux* » et « *actants technologiques* » (Lewis et Westlund, 2016), soulève des questions que la recherche sur les pratiques et l'épistémologie du journalisme commence à prendre en compte (Dagiral et Parasie, 2017).

Par exemple, l'enquête de responsabilité algorithmique (*algorithmic accountability reporting*, c'est-à-dire l'investigation journalistique sur les décisions algorithmiques, leurs biais potentiels et la manière dont elles façonnent la vision de la société par le public) n'a été introduite que récemment comme une nouvelle discipline pour les journalistes, et comme une contribution importante à la responsabilité publique en général (Diakopoulos, 2019).

Dans ce contexte, il apparaît nécessaire que l'évaluation de la performance d'un algorithme ne se limite pas à une mesure de son efficacité, mais incorpore également une évaluation de la conformité des décisions rendues vis-à-vis de l'intention des personnes l'ayant conçu. À cet égard, la transparence de l'algorithme permet en partie une telle évaluation. Par transparence, nous entendons la possibilité de comprendre l'intention de la personne ayant conçu l'algorithme et d'avoir accès aux critères utilisés pour aboutir à une décision (ainsi que leur pondération respective). Cet effort de transparence soulève des problèmes à la fois techniques et non techniques.

Du point de vue non technique, il est fréquent que des algorithmes d'apprentissage automatique doivent rester entièrement ou partiellement confidentiels pour des raisons de propriété intellectuelle. Du point de vue technique, identifier les critères de conception d'un algorithme par des outils de rétro-ingénierie[1] peut s'avérer extrêmement complexe (Sandvig, Hamilton et al., 2014 ; Datta, Tschantz et al., 2015 ; Kroll, Huey et al., 2017). De plus, le fait que ces algorithmes puissent faire évoluer leur modèle en étant continuellement entraînés avec de nouvelles données augmente encore la complexité de ce travail de rétro-ingénierie puisque le modèle à appréhender n'est pas figé dans le temps.

Bien que déjà difficile à garantir, cet idéal de transparence a pourtant ses limites (Ananny et Crawford, 2008). Dans le présent article, nous proposons de nous focaliser sur les limitations techniques qui font de la transparence une condition nécessaire, mais non suffisante, à la garantie d'une responsabilité algorithmique. Plus précisément, nous montrons que la robustesse face à des comportements « adversariaux » (*adversarial behavior*)[2] devrait également être prise en compte dans l'évaluation de la performance d'un algorithme. En effet, même si ce dernier est parfaitement transparent, c'est-à-dire si la rétro-ingénierie conclut que les critères intégrés dans un algorithme correspondent à son intention affichée, il se peut que des exemples adversariaux (*adversarial examples*) fassent dévier l'algorithme de son fonctionnement attendu (Goodfellow, McDaniel et al., 2018). Comme son nom l'indique, un exemple adversarial est une entrée, non observée

[1] C'est-à-dire l'étude d'un système existant dans le but de déterminer son fonctionnement et la manière dont il a été conçu.

[2] Il n'y a pas à l'heure actuelle de traduction largement adoptée en français du mot anglais *adversarial* tel qu'on le retrouve dans des expressions comme *adversarial example* ou *adversarial behavior*. Les traductions « contradictoire », « contrefactuel », « antagoniste » ou « adversarial » sont le plus souvent observées. Par souci de clarté, nous avons décidé de garder la traduction « adversarial » en français : elle permet notamment de préserver la proximité avec le terme anglais et ainsi éviter toute ambiguïté. Elle conserve également une référence explicite à la notion d'adversaire, centrale dans le domaine de la sécurité informatique. S'agissant d'une décision propre à notre article, nous avons entouré de guillemets les premières occurrences du mot « adversarial ». Nous les omettons dans la suite du document pour ne pas alourdir inutilement le texte.

lors de l'apprentissage du modèle, qui le fait dévier de ses spécifications publiques et est généré par un « adversaire » ayant un intérêt à déclencher une telle déviation.

Une caractéristique importante de ces exemples adversariaux est que leur exploitation ne demande pas de capacités importantes de la part de l'adversaire. En effet, s'agissant de données qui ne font qu'exploiter le manque de généricité d'un modèle, le fait qu'ils ne requièrent aucune modification du code de l'algorithme lui-même, ni même aucune intervention sur les articles qui sont utilisés pour entraîner l'algorithme en question. Cela signifie que même des personnes n'ayant pas participé à la conception d'un algorithme peuvent utiliser des exemples adversariaux pour le détourner de son objectif affiché. Les exemples adversariaux diffèrent en cela d'autres types d'attaques. Nous pouvons citer comme exemple l'empoisonnement de données (*data poisoning* : Biggio and Laskov 2012 ; Steinhardt, Pang et al. 2017) qui consiste précisément à modifier intentionnellement les données d'entraînement utilisées par un algorithme afin d'en biaiser le modèle.

En pratique, nous illustrons ce problème au moyen de deux cas d'application : d'une part un système de recommandation automatique d'articles, et d'autre part un détecteur automatique d'informations non fiables. Dans le premier cas, nous analysons un système de recommandation impliquant la classification automatique d'un corpus de plus de 4 000 articles issus du journal *Le Soir*, quotidien de référence en Belgique francophone, en diverses catégories génériques (International, National, Économie, Culture, Sports, Médias, Autre). Dans le second cas, nous analysons un système de détection d'informations douteuses à partir d'un corpus constitué, d'une part, d'articles issus de sites considérés comme non fiables ; et, d'autre part, d'articles issus du *New York Times* et du *Guardian*, tous couvrant approximativement les mêmes sujets et la même période. Nous montrons dans ces deux cas comment les modèles générés par des algorithmes de classification utilisés dans de tels systèmes peuvent être trompés par des modifications subtiles introduites dans les articles à analyser.

Après avoir passé en revue l'état de la littérature sur le sujet, nous détaillons la méthodologie que nous avons suivie dans les deux études de cas évoquées ci-dessus, tant pour l'entraînement initial des systèmes ciblés que pour la seconde phase consistant à concevoir des exemples adversariaux visant à exploiter le manque de généricité du modèle entraîné par l'algorithme. Nous explorons ensuite les conséquences concrètes que ces comportements adversariaux peuvent avoir dans le contexte d'une automatisation grandissante des processus de production d'information au sein des rédactions. Nous suggérons dès lors que la robustesse face aux comportements adversariaux, et plus généralement la sécurité de l'apprentissage automatique, soient prises en compte dans l'évaluation de la performance d'un algorithme. Enfin, nous abordons les différents défis que cette problématique soulève pour la pratique du journalisme et la nécessaire implication des journalistes comme garants de la pertinence des décisions automatiques.

État de l'art et travaux connexes

La recherche sur l'apprentissage automatique équitable et responsable (*fair and accountable machine learning*) peut être considérée comme la contrepartie technique de plusieurs axes de recherche non techniques qui ont mis en évidence le caractère critique des changements entraînés par l'automatisation, la collecte massive de données et la prise de décision basée sur des algorithmes, que ce soit pour les journalistes ou pour le public.

Par exemple, parmi les différentes approches d'une sociologie du journalisme computationnel, Anderson plaide en faveur d'une « *étude à orientation technologique du journalisme computationnel* », afin de discuter de la technologie dans les termes qui lui sont propres (Anderson 2012, p. 1016). Lewis et Usher soulignent également l'importance d'une approche du journalisme centrée sur la technologie et la nécessité de « *comprendre comment les idées, les pratiques et l'ethos des communautés de technologues pourraient être appliqués pour repenser les outils, la culture et le cadre normatif du journalisme lui-même* » (2013, p. 603). À cet égard, Lewis, Guzman et al. ont récemment « *abordé les questions ontologiques plus larges du journalisme automatisé* » à travers le prisme de la communication humain-machine, ce qui leur a permis de « *repositionner la technologie dans les processus sociaux du journalisme et de développer de nouvelles questions de recherche mieux adaptées à cette technologie* » (2020). Notre contribution se situe précisément dans ce type d'approches, en étudiant plus spécifiquement le cas concret des exemples adversariaux.

Les exemples adversariaux sont une des attaques possibles, répertoriées et investiguées dans le champ de recherche de la sécurité de l'apprentissage automatique (*machine learning security*), une branche du domaine de la sécurité informatique dont l'importance est croissante depuis quelques années au vu de l'essor de ces techniques dans de nombreux domaines d'application, parfois très sensibles (domaines médical, juridique ou militaire, par exemple). Barreno, Nelson et al. proposaient déjà une taxonomie de ces attaques en 2010, ensuite revue et mise à jour dans de nombreux travaux, comme ceux de Xue, Yuan et al. (2020).

L'attaque adversariale a d'abord été investiguée dans le champ de la classification d'images. Szegedy, Zaremba et al. montrent ainsi que des modifications imperceptibles par un humain introduites dans les images permettent d'induire des erreurs de prédiction des modèles d'apprentissage (2014). Goodfellow, Shlens et al. ont ensuite introduit la notion d'entraînement adversarial (2015), visant à améliorer les performances d'un modèle de classification d'images en le réentraînant avec des exemples adversariaux générés à partir des données d'entraînement initiales.

Si des perturbations imperceptibles sont faciles à générer dans le domaine des images, il n'en va pas de même dans le domaine du traitement automatique du langage (TAL) : contrairement à une matrice de pixels, un texte est issu d'un ensemble discret beaucoup plus restreint de symboles et la modification d'une séquence de texte n'est jamais entièrement imperceptible. De nombreux auteurs ont ainsi investigué comment définir et générer des exemples adversariaux pour du texte. Parmi eux, Ribeiro, Singh et al. introduisent la notion d'adversaire sémantiquement équivalent (*semantically equivalent adversary*) et des règles permettant d'en générer automatiquement (2018). Sato, Suzuki et al. se sont quant à eux demandé comment reconstruire un texte à partir de perturbations minimes réalisées dans l'espace vectoriel continu des plongements lexicaux (2018). Afin d'évaluer et de comparer les différentes tentatives d'attaques adversariales en TAL, Morris, Lifland et al. proposent un cadre théorique et des critères d'évaluation clairs pour ces exemples adversariaux : respect de la sémantique et de la grammaire ; degré de perceptibilité par un humain d'un changement donné ; par exemple (2020).

Bien qu'inspirée par ces travaux pour la génération et l'évaluation des exemples adversariaux présentés ci-dessus, notre contribution ne se situe pas dans l'amélioration

des systèmes développés dans ce champ de recherche. Elle vise davantage, dans une démarche interdisciplinaire, à importer ces résultats techniques dans le champ du traitement de l'information journalistique, à en déduire l'importance de la notion de robustesse dans l'évaluation des chaînes de traitement automatisé de l'information, et à identifier les implications de cet enjeu dans différentes thématiques de recherche liées au journalisme computationnel et introduites brièvement ci-dessous.

En ce qui concerne notre première étude de cas, nos résultats sont liés à la recherche analysant la personnalisation des contenus dans le contexte du journalisme numérique. Dans leur essai, Kunert et Thurman expliquent comment « *la personnalisation implicite qui déduit les préférences des données collectées* » gagne du terrain avec les appareils mobiles s'intégrant dans la routine quotidienne des individus (2019). Leur article révèle une tension entre la menace que représente la personnalisation pour la diversité de l'information, et les réactions positives des individus à des actualités sélectionnées automatiquement sur la base de leur comportement passé (Thurman, Moeller et al., 2019). Kunert et Thurman observent également que la personnalisation implicite est presque toujours mise en œuvre sans possibilité pour les individus de la désactiver. Ce dernier point est conforme aux observations antérieures de Kormelink et Costera Meijer selon lesquelles les individus ne sont pas disposés à gérer une trop grande complexité dans les paramètres de personnalisation (2014). Mais la combinaison de cette recommandation implicite avec les risques d'exemples adversariaux mis en avant dans le présent travail va à l'encontre de la volonté de contrôle des individus que mettent en avant leurs investigations. Plus généralement, nos résultats sont directement liés aux enjeux éthiques soulevés par les systèmes de recommandation, et en particulier à la question de l'opacité, abordée par Milano, Taddeo et al. (2020).

Notre seconde étude de cas s'ancre quant à elle dans l'analyse d'un phénomène qui a pris une ampleur particulière ces dernières années, à savoir la prolifération de la désinformation. La génération d'information fausse ou trompeuse, et leur détection par des opérations de vérification des faits (*fact checking*), se sont jusqu'ici le plus souvent déroulées manuellement (comme notamment discuté par Vargo, Guo et al., 2018). Mais les progrès récents en traitement automatique du langage ont amplifié le risque de « *désinformation neuronale* » (*neural fake news*), c'est-à-dire de contenu faux ou trompeur généré par des réseaux de neurones artificiels (Zellers, Holtzman et al., 2019). La quantité massive d'articles que des outils tels que GPT-3 (Brown, Mann et al., 2020) peuvent générer rend impraticable la vérification manuelle des informations et pose dès lors la question de leur détection automatisée. De nombreuses contributions techniques ont été faites dans ce domaine, notamment recensées par Zhou et Zafanari (2020). Parmi celles-ci, nous nous intéressons en particulier aux systèmes évaluant la fiabilité d'une information en se basant exclusivement sur le contenu du texte (à la fois sémantique et syntaxique), sans faire appel à des outils de vérification externe de l'information contenue dans l'article. Néanmoins, l'ensemble de cette problématique technique se heurte dans sa formulation à la difficulté inhérente au processus de définition même de ce qu'est la désinformation ou des critères requis pour considérer une information comme fiable (Tandoc, Lim et al., 2018). Alors que le caractère sensible de cette thématique exige des garanties accrues de responsabilité algorithmique, on peut constater que cette difficulté à définir ce qu'est une information fiable accroît la probabilité que des informations fausses ou trompeuses générées par des adversaires soient classées comme vraies. Notre

contribution s'inscrit dès lors dans une volonté de mettre en avant le risque sous-estimé des comportements adversariaux dans le cadre d'un déploiement de ces systèmes de détection d'informations non fiables et d'appuyer l'analyse faite par Lazer, Baum et al. d'une nécessaire approche multidisciplinaire de la lutte contre la désinformation (2018).

Notre contribution est également ancrée dans le champ de recherche plus large analysant l'impact de l'intelligence artificielle dans différents domaines, et en particulier dans le journalisme. Nos résultats peuvent ainsi être considérés comme une contribution à la clarification de ce que l'intelligence artificielle peut et ne peut pas réaliser, qui est l'une des questions abordées par Broussard (2018). De manière plus spécifique au champ journalistique, un ouvrage de Broussard, Diakopoulos et al. (2019) examine plusieurs questions qui recoupent les nôtres. Nous pointons en particulier celle de la nécessaire porosité entre les différentes disciplines impliquées, tant pour l'utilisation au jour le jour de ces nouveaux outils, que pour leur conception et la recherche qui y est associée.

Du point de vue de l'utilisation, nous montrons en effet que même si la personne ayant conçu un algorithme y intègre certaines valeurs, un adversaire peut être en mesure d'abuser de cet algorithme pour imposer des valeurs contradictoires et un comportement différent de celui attendu. Ce constat appelle les journalistes à jouer encore davantage un rôle d'interface entre les technologies basées sur les données et les usagers finaux des applications d'information. Du point de vue de l'importance d'une recherche interdisciplinaire mentionnée ci-dessus, nous la confirmons en adaptant une méthodologie classique du domaine de la sécurité de l'information au contexte du journalisme numérique (voir le détail de cette approche au début de la section suivante consacrée à la méthodologie).

Enfin, les limites de l'idéal de transparence et son implication dans la notion de responsabilité algorithmique sont discutées en profondeur par Ananny et Crawford (2018). Ils énumèrent dix lacunes de cet idéal, parmi lesquelles les préoccupations d'ordre technique sont étroitement liées à la question de la sécurité de l'apprentissage automatique. Les auteurs mettent ainsi l'accent sur « *les systèmes que même les personnes les ayant conçus ne sont pas en mesure de comprendre, à cause de leur ampleur et de la vitesse à laquelle ils sont conçus* » (Burrel, 2016 ; Crain, 2018). Nous élargissons ce point de vue en montrant que des exemples adversariaux peuvent encore aggraver cette situation et qu'il n'existe pas à ce jour de cadre théorique permettant de garantir une robustesse à ce type d'attaque. En outre, dans le cas d'exemples adversariaux, la transparence peut même être préjudiciable à la responsabilité algorithmique en facilitant la tâche consistant à trouver des modifications mineures d'un article conduisant à une classification erronée (Kroll, Huey et al., 2017). Ananny et Crawford (2018) concluent en suggérant une typologie alternative de la gouvernance algorithmique (dans laquelle sont reconnues les limites de la transparence). Nous souscrivons à cette démarche en appelant à davantage considérer les algorithmes d'apprentissage automatique comme des aides à la décision plutôt que comme des décideurs autonomes à qui il est difficile de faire rendre des comptes.

Hypothèse de recherche et méthodologie

D'un point de vue méthodologique, nous nous inspirons de l'approche standard en cryptographie moderne consistant à affiner de manière itérative une définition en

raisonnant à partir de contre-exemples (Katz et Lindell, 2014). Ainsi, nous nous intéressons à une définition opérationnelle des performances d'un algorithme qui puisse inclure le concept de responsabilité algorithmique, tel que discuté par exemple par Diakopoulos (2015). Nous considérons en effet que spécifier les objectifs d'un algorithme en termes précis et opérationnels est utile à la fois pour les personnes l'ayant conçu (de manière à ce qu'elles sachent ce qu'elles visent à concevoir) et pour personnes qui l'utilisent (de manière à ce qu'elles sachent ce qu'elles peuvent en attendre). Sur cette base, notre hypothèse de recherche est qu'une définition de la performance d'un algorithme responsable se doit d'incorporer la notion de robustesse à l'égard de comportements adversariaux ; et en particulier que la responsabilité algorithmique ne peut pas être réduite à une question de transparence.

Les exemples auxquels nous nous intéressons pour étayer les limites techniques de la transparence et la nécessaire prise en compte de la robustesse à l'égard de comportements adversariaux dans l'évaluation de la performance d'un algorithme sont deux systèmes de classification automatique d'articles basée sur leur contenu. Dans le premier cas, la classification vise à attribuer automatiquement une catégorie à l'article analysé. Cette opération de catégorisation peut être considérée comme un composant d'un système plus large visant à recommander des articles d'un certain type sur base des préférences de l'usager. Dans le second cas, la classification des articles est une classification binaire visant à déterminer si l'article en question est une information digne de confiance.

Notons que nous ne prétendons pas à l'optimalité des systèmes présentés ici et notre contribution ne se situe pas dans une potentielle amélioration de systèmes actuellement utilisés dans des rédactions. Pour autant, les algorithmes de classification utilisés correspondent à l'état de l'art dans le domaine et représentent à cet égard des outils simples mais vraisemblables que l'on pourrait trouver dans des processus de traitement de l'information. Notre unique ambition est de montrer que la robustesse vis-à-vis de comportements adversariaux devrait être prise en compte dans l'évaluation de la performance des algorithmes (utilisés ici dans un contexte journalistique) et que la transparence de ce type de systèmes est insuffisante pour garantir la pertinence des décisions prises. Le fait que les systèmes étudiés ne soient pas nécessairement optimaux ne déforce pas la validité de nos conclusions dans la mesure où ces dernières mettent uniquement en évidence l'incomplétude de la définition de responsabilité algorithmique réduite à une question de transparence. À cet égard, nous insistons sur le fait que, formellement, invalider une définition ne nécessite pas de montrer son incomplétude dans de nombreux cas. Au contraire, l'objectif d'une définition est d'être générale et indépendante des cas applicatifs. Ainsi, la seule existence d'un contre-exemple, c'est-à-dire ici d'un système de classification basé sur des techniques couramment utilisées, mais faillible face à des comportements adversariaux, confirme cette incomplétude.

Description des cas applicatifs

Dans cette section, nous exposons les modalités de collecte de nos données et la succession des traitements algorithmiques qui leur est appliquée en vue de leur classification.

Système de recommandation d'articles

Nous envisageons ici un système de recommandation d'articles dans lequel un ensemble de catégories est prédéfini par les journalistes et dans lequel les articles ingérés se voient attribuer automatiquement l'une d'elles. Ensuite, en fonction des préférences de l'usager et des catégories attribuées automatiquement, un score de préférence est calculé pour chaque article. Un tel score peut être utilisé pour hiérarchiser les articles présentés à l'usager dans son fil d'actualité. Dans le contexte de la présente étude, et pour éviter d'avoir à gérer une problématique de manipulation de données personnelles que ce type d'application soulève, nous nous sommes concentrés sur la qualité et la robustesse de l'étape de classification automatique d'articles.

Concrètement, nous avons utilisé une base de données d'un peu plus de 4 000 articles issus du journal *Le Soir*, quotidien de référence en Belgique francophone. Chaque article de cette base de données inclut un titre, le corps de l'article, et une catégorie attribuée par la rédaction du *Soir*, parmi les sept possibilités suivantes : International, National, Économie, Culture, Sports, Médias, Autre. Ces catégories sont mutuellement exclusives (un article appartient à une et une seule catégorie). Les articles ont été collectés sur une période de six mois, entre août 2018 et janvier 2019, au moyen d'un script d'indexation (*web crawling*) développé spécifiquement pour les besoins du projet. Comme indiqué dans la Table 1 ci-dessous, le nombre d'articles par catégorie varie entre 500 et 800 (excepté pour la catégorie Médias qui ne contient que 104 articles).

Catégorie	**Nombre d'articles**	**Proportion**
International	772	19 %
National	772	19 %
Économie	519	13 %
Culture	580	14 %
Sports	825	20 %
Médias	104	3 %
Autres	561	14 %
Total	**4133**	**100 %**

Table 1. Nombre d'articles par catégorie

Une fois la base de données assemblée, celle-ci a été séparée en une partie DB_{train} utilisée pour l'entraînement du classifieur et une autre DB_{test} pour tester la qualité du classifieur. Les catégories assignées manuellement par les journalistes ont été utilisées à la fois pour l'entraînement du classifieur (dans le cadre d'un apprentissage automatique supervisé), et pour l'évaluation de la qualité du classifieur (en comparant pour un article donné la prédiction automatique du classifieur à la catégorie assignée manuellement). Par ailleurs, afin d'estimer de manière fiable la qualité du classifieur, nous avons utilisé un système de quintuple validation croisée (*k-fold cross validation*, avec $k = 5$). Ainsi, l'évaluation de la qualité de chaque algorithme de classification que nous avons testé résulte de l'agrégation de cinq évaluations distinctes, chacune de ces évaluations étant réalisée en utilisant

quatre cinquièmes des données pour DB_{train}, puis le cinquième restant pour DB_{test}, et en bouclant ainsi sur chacun des cinquièmes de la base de données complète.

Le processus de classification des articles mis au point dans le cadre de cette recherche est le suivant. Il débute avec une série d'opérations classiques en traitement automatique du langage : retrait de la ponctuation et racinisation (*word stemming*) afin d'associer l'ensemble des variations d'un mot à sa racine. Afin de représenter l'ensemble des mots d'un article dans un format adapté au traitement automatique, nous avons ensuite comparé différentes techniques. La plus simple est la méthode consistant à utiliser un dictionnaire ne contenant que les 20 000 mots les plus fréquemment rencontrés dans notre base de données et à représenter chaque article comme un « sac de mots » (*bag of words*) constitués d'occurrences de ces mots les plus fréquents. Une alternative à cette méthode simple est de constituer un dictionnaire contenant non pas les mots les plus fréquents mais les mots les plus saillants de chaque article à l'aide de l'indice TF/IDF, pour *term frequency / inverse document frequency* (Spärck Jones, 1972). Cet indice est le rapport entre la fréquence d'un terme dans un document donné (ici : un article) et le nombre de documents dans lequel ce terme apparaît. Il est apparu néanmoins que ces deux options (basées sur la fréquence et la saillance) donnaient des résultats similaires et nous nous sommes donc focalisés sur la méthode des mots les plus fréquents. Enfin, dans une approche plus élaborée, nous avons utilisé la technique des plongements lexicaux (*word embeddings*) et plus spécifiquement le modèle *Word2Vec* (Mikolov, Chen et al., 2013) qui permet la représentation de mots dans un espace vectoriel continu qui préserve certaines relations sémantiques entre les mots.

Basé sur ces représentations des articles, nous avons ensuite comparé les performances de trois algorithmes de classification utilisant l'apprentissage automatique : le classifieur naïf bayésien (*naïve Bayes*, NB), le perceptron multicouche (*multi-layer perceptron*, MLP), et un réseau de neurones artificiels récurrent (*recurrent neural network*, RNN).

NB est un classifieur probabiliste simple basé sur le théorème de Bayes. Le MLP (Rumelhart, Hinton et al., 1986) est un classifieur un peu plus complexe constitué de plusieurs classifieurs simples. C'est un exemple basique de réseau de neurones artificiels fréquemment utilisés pour ce type de tâches. Enfin, le RNN est un type de réseau de neurones artificiels dit « profond » (*deep neural network*) similaire au MLP, mais faisant appel à un nombre accru de couches de classifieurs. Contrairement au MLP, le RNN présente l'avantage de tenir compte dans son processus de classification du contexte de chaque mot et de leur ordre. Différentes variantes de RNN ont été utilisées et combinées, parmi lesquelles le RNN bidirectionnel (Schuster et Paliwal, 1997) et le RNN de mémoire longue à court terme (*Long Short-Term Memory*) proposé par Hochreiter, Sepp et al. (1997). Les détails techniques à propos de ces algorithmes ainsi que les paramètres utilisés ne sont pas nécessaires à la compréhension de nos résultats et nous invitons dès lors le lecteur intéressé à se référer à notre article traitant spécifiquement de cette étude de cas (Descampe, Massart et al., 2021). Notons que les techniques référencées ci-dessus, bien que basées sur des algorithmes développés il y a plus de 20 ans, sont toujours largement utilisées aujourd'hui et ont vu leurs performances augmenter significativement, notamment grâce à la disponibilité de plus grandes quantités de données et d'une puissance de calcul accrue.

En pratique, nous avons entraîné les classifieurs NB et MLP avec les séquences de taille fixe fournies par la technique des « sacs de mots », tandis que le RNN a été entraîné avec les représentations de taille variable fournies par les plongements lexicaux. Nous avons évalué l'exactitude (*accuracy*) obtenue par ces trois combinaisons d'outils, c'est-à-dire la proportion d'articles correctement classés par rapport au nombre total d'articles[3]. Une performance de 85 % a été obtenue pour les classifieurs NB et MLP et de 80 % pour le RNN. Cette différence peut s'expliquer par le fait que l'identification de la catégorie d'un article est une tâche relativement simple, à la portée de techniques peu complexes, tandis que le RNN est en mesure d'appréhender des caractéristiques du texte plus subtiles, mais nécessite de manière générale une plus grande quantité de données d'entraînement pour mener à de bons résultats.

En tout état de cause, ce qu'il est important de noter dans le cadre de notre étude est que ces trois classifieurs atteignent une exactitude largement supérieure à un choix aléatoire parmi les sept catégories. Au vu de ces performances, nous pouvons conclure que les outils d'apprentissage automatique proposés satisfont aux critères d'une responsabilité algorithmique ramenée à une exigence de transparence : lorsque des algorithmes de classification bien connus du type NB, MLP ou RNN, entraînés et paramétrés de manière transparente et conforme à leur objectif affiché sont utilisés dans un environnement contrôlé, ils se comportent de la manière attendue et parviennent à classer des articles presque aussi bien que le feraient manuellement des journalistes.

Détecteur de désinformation

Dans cette seconde étude de cas, nous avons également analysé un exemple de classification automatique d'articles, visant cette fois à déterminer si un article donné était fiable ou relevait possiblement de la désinformation. Une différence importante par rapport au premier cas, outre qu'il s'agit ici d'une classification binaire et non en sept catégories différentes, est la difficulté inhérente à la définition même de ce qu'est la désinformation et à la collecte d'un ensemble de données permettant de construire un classifieur fiable. Dans le cadre de cette étude, nous avons recherché des corpus existants tels que celui fréquemment utilisé proposé par Wang (2017). Malheureusement, les articles y sont souvent très courts et ne permettaient dès lors pas d'investiguer de manière pertinente les comportements adversariaux (qui impliquent de faire des modifications subtiles du texte en vue d'un changement de classification). Nous avons finalement construit notre propre corpus en combinant deux ensembles de données.

Nous avons d'une part utilisé un ensemble de données publiques de 3 500 articles issus de Kaggle[4] collectés à partir de 244 sites web considérés comme non fiables sur une période de 30 jours autour des élections américaines de 2016. D'autre part, nous avons construit un corpus du même nombre d'articles, et couvrant la même période, issus de deux sources

[3] Les mesures les plus appropriées pour analyser en détail les performances d'un classifieur dont les classes ne sont pas parfaitement équilibrées (voir Table 1) sont la précision, le rappel et la *f-mesure*. Néanmoins, nous avons opté pour la simplicité de l'exactitude (*accuracy*), car elle est suffisante pour atteindre les objectifs de l'analyse réalisée ici, à savoir vérifier que le classifieur fonctionne suffisamment bien et qu'il est ensuite possible de faire changer de classe des articles par des modifications légères (voir la section suivante consacrée aux exemples adversariaux).

[4] Kaggle est une plateforme hébergeant des projets de recherches scientifiques réalisées à partir de corpus de données et de code disponibles en ligne. Le corpus utilisé dans cet article est accessible à www.kaggle.com/mrisdal/fake-news.

considérées comme fiables (le *New York Times* et le *Guardian*) sur des sujets internationaux (*World News*) ou américains (*US News*).

Afin de nous assurer que les sujets traités dans les articles ne soient pas trop différents entre les deux ensembles et qu'ils ne puissent dès lors pas être utilisés pour en identifier le caractère fiable ou non fiable, nous avons utilisé la méthode t-SNE (*t-distributed stochastic neighbour embedding*) développée par van der Maaten et Hinton (2008). Elle permet de visualiser des objets appartenant à des espaces de grande dimension (comme le sont les articles de notre corpus) dans un espace à deux dimensions d'une manière telle qu'avec une grande probabilité les objets similaires soient modélisés par des points proches et les objets fort différents soient modélisés par des points éloignés. Concrètement, la t-SNE a été réalisée à partir des vecteurs de très grande dimension représentant chaque article et obtenus avec l'indice TF-IDF, et dont la dimension a été réduite par décomposition en valeurs singulières (*truncated-SVD*). La Figure 1 présente le résultat de cette analyse et indique que la distribution des sujets est similaire entre les articles fiables et non fiables (il n'y a pas de regroupement clair des articles de différents types), alors que par exemple on observe une séparation nette des sujets internationaux et américains dans le sous-corpus des articles fiables (voir figure de droite). Cette analyse suggère que les sujets traités dans les articles ne contribuent pas de manière évidente à la discrimination entre les articles fiables et non fiables. Bien que nous ne puissions pas en avoir une garantie formelle, nous avons considéré dans la suite de ce travail que la base de données ainsi construite était d'une qualité suffisante pour contribuer à répondre de manière pertinente à notre question de recherche.

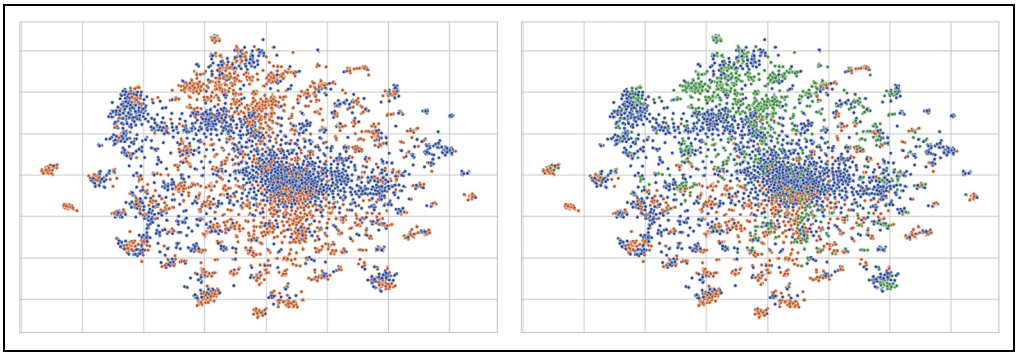

Figure 1. Visualisation des sujets traités au sein du corpus d'articles fiables et non fiables au moyen de l'outil t-SNE. Gauche : articles non fiables (points bleus) et fiables (points orangés). Droite : articles non fiables (points bleus), fiables « international » (points orangés) et fiables « États-Unis » (points verts)

Une fois notre base de données constituée, nous avons adopté sensiblement la même approche que dans la première étude de cas, à savoir un prétraitement classique en TAL (suppression de la ponctuation, extraction des racines de mots), une vectorisation basée sur l'indice TF/IDF ou sur les plongements lexicaux (*Word2Vec*), et une classification au moyen des algorithmes suivants : une simple régression logistique[5] et un RNN. Afin

[5] Nous avons également testé la classification bayésienne naïve comme dans le cas précédent, qui aboutit à des résultats similaires.

d'évaluer la qualité de notre système, nous avons, comme dans le cas précédent, procédé à une quintuple validation croisée.

L'exactitude (*accuracy*) obtenue dans cette seconde étude de cas dépasse 90 % pour la combinaison de l'outil TF/IDF suivi d'une régression logistique. Elle est légèrement inférieure pour la combinaison des plongements lexicaux avec le RNN (~85 %), mais l'analyse faite en fonction du nombre d'articles tend à indiquer qu'un nombre plus élevé d'entre eux permettrait d'atteindre une plus grande exactitude. Comme pour la catégorisation automatique d'articles, nous n'entrons pas ici dans les détails techniques du paramétrage de ces algorithmes et renvoyons le lecteur intéressé à notre article traitant spécifiquement de cette étude de cas (Descampe, Massart et al., 2021).

Sur la base des résultats obtenus, nous constatons à nouveau que l'outil de classification développé ici dans le contexte de la détection d'articles faux ou trompeurs satisfait aux critères d'une responsabilité algorithmique ramenée à une question de transparence. Utilisé dans des conditions « normales », c'est-à-dire avec des données de test qui n'ont pas été spécifiquement façonnées pour tromper l'algorithme, le système de classification se comporte de manière attendue et parvient dans une très grande majorité des cas à distinguer les informations fiables des informations non fiables.

Exemples adversariaux

Dans cette section, nous partons des deux classifieurs décrits ci-dessus et qui fonctionnent correctement dans des conditions normales d'utilisation. Sur cette base, nous investiguons de quelle manière nous pouvons construire des données de test (c'est-à-dire qui n'ont pas été utilisées pour leur entraînement) qui fassent dévier ces classifieurs de leur comportement attendu : soit en attribuant à l'article une autre catégorie que celle initialement assignée par le classifieur (et correspondant à la catégorie assignée manuellement par le ou la journaliste), soit en classant l'article comme fiable alors qu'il s'agit d'un article non fiable.

Intuition

Avant de présenter les résultats expérimentaux concrets obtenus dans chaque cas, nous proposons d'abord une compréhension intuitive de la manière dont des exemples adversariaux fonctionnent pour provoquer une classification erronée.

La figure 2 représente un classifieur élaboré à partir d'un apprentissage supervisé basé sur 12 articles d'entraînement : six articles considérés comme fiables (F) et six considérés comme non fiables (NF). Le classifieur qui en résulte est symbolisé par la ligne noire qui distingue correctement les données d'entraînement. Il y a également un article-test t_a, inconnu du classifieur, qui est initialement (et correctement) classé comme non fiable. L'objectif poursuivi par un adversaire sera de modifier t_a de manière minimale en un article similaire t_a' de telle sorte que t_a' soit classé incorrectement comme fiable.

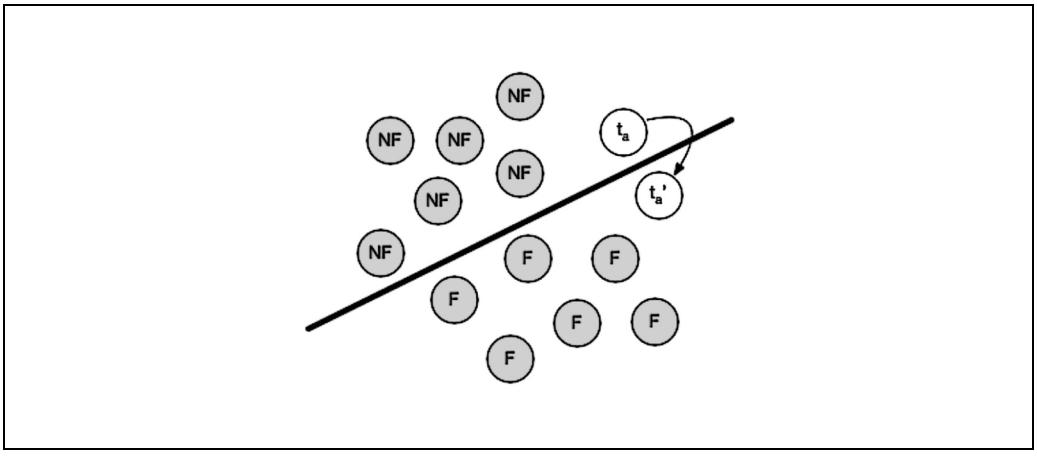

Figure 2. Un exemple adversarial. NF = article non fiable. F = article fiable. t_a = article test (non fiable). t_a' = exemple adversarial, similaire à t_a mais classé comme fiable.

La difficulté avec laquelle il sera possible de construire un exemple adversarial dépend de la distance statistique entre la catégorie initiale et la catégorie cible. Dans notre exemple, il s'agit de la distance de la classe non fiable à la classe fiable, mais cela vaut aussi pour notre première étude de cas avec les sept catégories différentes d'articles. Intuitivement, on comprend, par exemple, qu'il sera probablement plus facile de « transférer » un article de la catégorie Culture vers la catégorie Médias que de Culture vers Économie.

Un autre point qu'il nous semble important de souligner est que l'objectif d'un exemple adversarial n'est pas de modifier un article arbitrairement jusqu'à ce qu'il soit classé tel que le souhaite l'adversaire. L'objectif est bien de trouver les modifications minimales permettant que la perception du contenu et sa compréhension sémantique soient les moins altérées possible pour un humain. Un exemple classique illustrant ce principe est issu du traitement d'image : Eykholt, Evtimov et al. (2018) ont montré qu'il était possible de modifier légèrement des signaux routiers de manière à ce qu'ils soient interprétés erronément par une voiture autonome (avec les conséquences que l'on peut imaginer), mais toujours correctement par un conducteur humain.

Fabrication d'exemples adversariaux

Dans les deux cas envisagés ici, la méthodologie de fabrication d'exemples adversariaux est similaire et consiste à identifier les mots contribuant le plus au classement d'un article dans une classe donnée. Pour ce faire, un score est calculé pour chaque mot d'un article et pour chaque classe, indiquant la manière dont ce mot « contribue » à la probabilité que l'article qui le contient appartienne à une classe donnée. Le calcul de ce score peut être réalisé en se plaçant dans deux modes opératoires distincts : un mode « boîte noire » où l'on considère que ne sont connues que les entrées et sorties du classifieur et un mode « boîte blanche » où l'on a accès aux paramètres internes du classifieur. Dans ce deuxième cas, une connaissance plus fine de la contribution de chaque mot est possible grâce à l'application de méthodes classiques de descente de gradient, comme proposé par exemple par Liang, Li et al. (2018). Ces deux modes correspondent à des scénarios distincts d'attaque sur un système automatisé, le mode « boîte noire » étant celui

nécessitant le plus faible degré d'intrusion. Ces deux modes ont été investigués dans le cadre de notre étude mais le choix d'un mode ou de l'autre ne modifie pas la teneur des conclusions de notre analyse et ne sera dès lors pas détaillé plus avant dans cet article.

Une fois ce score calculé pour chaque mot, il reste à décider quels mots retirer ou ajouter afin de faire basculer la classification vers la classe cible. Le choix des mots à modifier dépendra bien évidemment à la fois de la manière dont ils influencent la classification, mais également de leur poids sémantique dans le contenu de l'article : seront ainsi privilégiés les mots qui n'altèrent pas ou peu le contenu sémantique de l'article. Si une fabrication manuelle de ces exemples adversariaux paraît la méthode la plus immédiatement accessible (voir exemples dans les sous-sections suivantes), il existe également des perspectives d'automatisation qui seront brièvement explorées en fin de section.

Exemple de corruption du système de recommandation d'articles

Pour notre première étude de cas, nous présentons ci-dessous un exemple classé initialement correctement comme national par les deux classifieurs NB et MLP décrits plus haut. Pour cet article, la modification en gras est suffisante pour que le classifieur NB classe erronément l'article dans la catégorie Économie.

> Flandre : des mesures préventives contre les attaques de loups
>
> […] Des investissements passés peuvent être subsidiés s'ils remplissent les conditions. ~~Ce~~ **Normalement, ce** nouveau règlement doit être prêt pour début avril. Les éleveurs qui équipent leur terrain d'un enclos contre les loups pourront récupérer 80 % de leurs investissements. […]

Cet exemple a été trouvé en cherchant manuellement des articles ne nécessitant que de légères perturbations et pour lesquels les mots les plus impactants étaient perçus comme neutres. Malgré l'identification quelque peu artificielle de cet exemple, cela illustre néanmoins la faisabilité technique d'une déviation intentionnelle et difficile à percevoir par un humain de notre système de recommandation.

Pour compléter l'exemple ci-dessus identifié manuellement – et puisque, contrairement à l'autre étude de cas, le système a ici recours à plus de deux catégories d'articles – il nous a paru pertinent d'estimer la perturbation moyenne nécessaire à un article pour le faire atterrir dans une des six autres classes cibles. Pour ce faire, nous avons utilisé comme mesure de distance entre un article et une classe donnée la probabilité que cet article appartienne à la classe en question. Ensuite, en utilisant le score des mots décrit ci-dessus, nous avons calculé le nombre de fois que les mots avec un score élevé devaient être retirés/ajoutés et avons utilisé ce nombre comme une mesure de la perturbation nécessaire pour que l'article atteigne chacune des autres classes cibles.

À titre d'exemple, une perturbation de 5 % dans un article de 100 mots signifie que nous avons dû ajouter 5 fois le mot avec le score le plus élevé de la classe cible. La figure 3 présente le résultat de cette analyse sous la forme d'un diagramme en boîte à moustache pour le classifieur NB, pondéré par la moyenne de l'ensemble des articles. Le classifieur RNN mène à des résultats similaires.

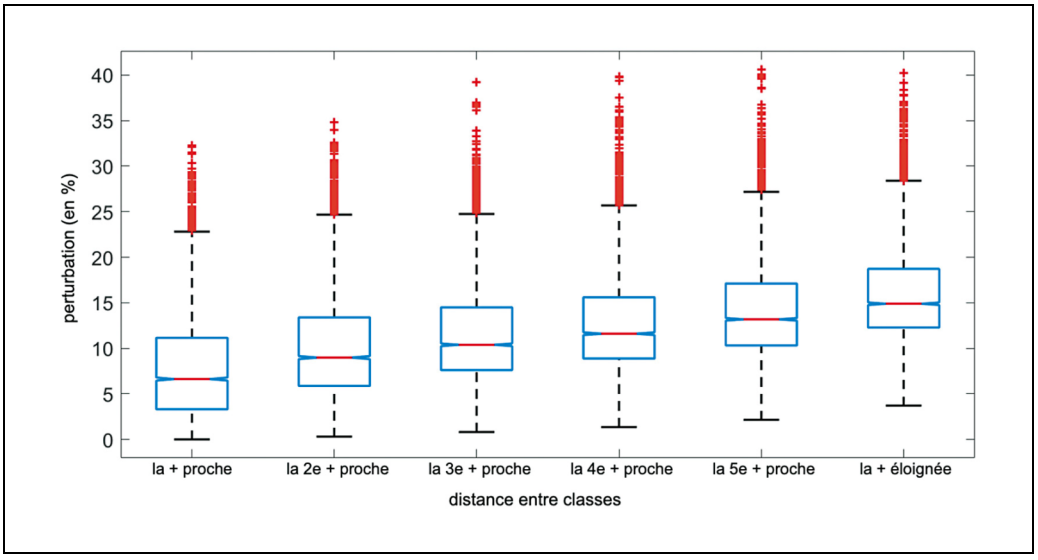

Figure 3. Perturbation minimale requise pour qu'un article donné atteigne une classe cible, en fonction de la distance entre l'article et cette classe.

L'observation principale que nous pouvons faire est que la médiane de chaque boîte à moustache se situe entre 6 % (pour la classe la plus proche) et 15 % (pour la classe la plus éloignée). Cela signifie donc qu'une perturbation minimale d'environ 6 % pour la classe la plus proche et jusqu'à 15 % pour la classe la plus éloignée est nécessaire pour classer erronément 50 % des articles. Nous observons également que pour certains articles, des perturbations beaucoup plus faibles sont suffisantes : inférieure à 1 % pour la classe la plus proche, et 3 % pour la classe la plus éloignée.

Bien sûr, cette analyse purement quantitative de la fabrication d'exemples adversariaux est très limitée sur le plan opérationnel. En particulier, elle ne tient pas compte des contraintes de cohérence sémantique, ni du fait qu'un grand nombre de modifications de mots tout à fait neutres (comme dans l'exemple présenté ci-dessus) peut mener à des exemples adversariaux plus difficiles à détecter qu'avec quelques modifications de mots sémantiquement associés à une catégorie spécifique. Cela donne néanmoins une indication de la faible ampleur des changements à réaliser et souligne à nouveau la fragilité de ces systèmes de classification automatisés.

Exemple de corruption du détecteur de désinformation

En suivant la même méthodologie pour notre deuxième étude de cas, nous avons constitué une liste de mots contribuant le plus à la classification d'un article comme fiable et une liste de ceux contribuant le plus à le rendre non fiable. Nous avons ensuite manuellement remplacé certains des mots de la deuxième liste par ceux de la première d'une manière préservant au maximum la sémantique de l'article. Sur base des listes ainsi établies, une observation intéressante est que les modifications faites pour la régression logistique et celles pour le RNN, bien que non exactement identiques, sont très similaires. Ce dernier point tend à confirmer ce que Tramèr, Papernot et al. ont mis en lumière

(2017), à savoir une transférabilité élevée des exemples adversariaux entre différents modèles d'apprentissage automatique.

L'exemple ci-dessous est initialement détecté comme non fiable avec une probabilité de 90 % par la régression logistique et de 65 % par le RNN.

> *In what is being described as another "bizarre" attempt to sabotage her own campaign, Hillary Clinton has desecrated a series of beloved US symbols, including punching a bison, setting fire to the Stars & Stripes and spitting at Jerry Seinfield.* […] *Having already become the unwitting focus of various health scares and FBI investigations, Mrs. Clinton's campaign is as orderly as a Marx Brothers movie.* […] *Hillary's erratic behaviour has seen her sing the Star-Spangled Banner in Korean, dress as Oprah Winfrey for Halloween and pebble-dash Mount Rushmore. Remarked a flummoxed advisor: "She keeps doing the unthinkable – like making Donald Trump electable".*

À partir des listes de mots discriminants, nous avons été en mesure de fabriquer manuellement des exemples adversariaux. Ainsi, la modification suivante a permis de faire passer la probabilité de non fiabilité obtenue par le RNN de 65 % à 45 %, faisant ainsi basculer l'article du côté fiable.

> […] *attempt to sabotage her own campaign, ~~Hillary~~* ***Mrs*** *Clinton has desecrated*
>
> […]
>
> *Mrs. Clinton's campaign is as ~~orderly~~* ***neat*** *as a Marx Brothers movie* […]

Dans le cas de la régression logistique, ces changements font passer la probabilité de non fiabilité de 90 % à 55 % (insuffisant donc), mais en changeant la deuxième modification de la manière indiquée ci-dessous, la probabilité passe sous les 50 % et l'article est dès lors classé comme fiable.

> […] *attempt to sabotage her own campaign, ~~Hillary~~* ***Mrs*** *Clinton has desecrated*
>
> […]
>
> *a flummoxed ~~advisor~~* ***minister****: 'She keeps doing the unthinkable* […]

S'agissant d'une opération manuelle, nous n'avons pas répété ce processus sur un grand nombre d'articles. Néanmoins, sur les cinq articles non fiables que nous avons traités, nous sommes parvenus à forcer une classification erronée en changeant un maximum de 15 mots. Ces cinq exemples ne sont évidemment pas statistiquement significatifs. Ils constituent néanmoins des contre-exemples confirmant que la définition de responsabilité algorithmique ne peut être réduite à une question de transparence et que la performance d'un algorithme devrait être également évaluée sur la base de sa robustesse à des comportements adversariaux.

En comparaison à notre première étude de cas où les modifications adversariales devaient se focaliser principalement sur des mots sémantiquement neutres afin de minimiser la probabilité d'une perception humaine, nous observons que dans le cas de la détection de

la désinformation, les modifications peuvent s'opérer sur des mots davantage marqués sémantiquement (en remplaçant « Hillary » par « Mrs. » par exemple), ce qui facilite potentiellement la tâche d'un adversaire. Cela peut s'expliquer par le fait que la nature fiable ou non fiable d'un article est plus difficile à caractériser que les catégories identifiées dans notre système de recommandation d'articles.

Essai de génération automatique

Si une fabrication manuelle de ces exemples adversariaux paraît la méthode la plus immédiatement accessible (comme nous le décrivons ci-dessus), nous avons également brièvement investigué les perspectives d'automatisation dans le contexte de la détection de la désinformation. Nous avons ainsi évalué une première approche basique consistant à remplacer les mots indicateurs d'une non fiabilité par des synonymes issus d'un dictionnaire. Un exemple est donné ci-dessous :

> […] *symbols, including punching a* ~~*bison*~~ ***buffalo*** […] *a group of* ~~*Girls*~~ ***Woman*** *guides*
> […] *various health scares and FBI* ~~*investigations*~~ ***inquiry****, Mrs Clinton's campaign* […]

Sur les 100 articles testés, nous sommes parvenus à modifier la classification de 22 % d'articles classés avec le RNN et 32 % avec la régression logistique. Le taux plus faible obtenu avec le RNN peut s'expliquer par le fait que notre approche est davantage en mesure de tromper des modèles ne tenant pas compte de l'ordre et du contexte des mots.

La possibilité d'automatiser au moins partiellement la fabrication d'exemples adversariaux, combinée au développement récent de la production automatique de texte, est évidemment une perspective assez sombre en matière de désinformation et constitue à cet égard un champ de recherche qu'il serait intéressant d'approfondir.

Vers une responsabilité algorithmique robuste

Les deux cas applicatifs présentés dans le présent article mettent en avant le fait qu'une définition de la responsabilité algorithmique qui soit uniquement focalisée sur la transparence ne prend pas en compte certains risques inhérents à l'automatisation et que la robustesse face à des comportements adversariaux devrait être prise en compte dans l'évaluation de la performance d'un algorithme. Dans cette section, nous discutons les observations réalisées sur la base de nos deux études de cas dans le contexte plus large de l'automatisation de la production de l'information, tel qu'analysé notamment par Thurman, Lewis et al. (2019).

Observons par exemple l'impact que de tels comportements pourraient avoir dans la personnalisation du contenu, qui est une application directe de notre première étude de cas. Comme discuté par Bodò (2019), les objectifs et la mise en œuvre de la personnalisation peuvent varier, mais elle n'en reste pas moins une tendance lourde observée au cours de la dernière décennie. Dans ce contexte, Helberger (2019) s'interroge sur la menace que représentent les systèmes automatisés de recommandation et de personnalisation d'information au niveau du rôle que jouent les médias dans une société démocratique. Un élément central de sa contribution est que les valeurs qu'il s'agit d'optimiser dans un tel système automatisé dépendent en grande partie du modèle démocratique considéré. On peut constater néanmoins que, quel que soit ce modèle, les

comportements adversariaux peuvent avoir un impact significatif. Par exemple, dans une perspective libérale, l'indépendance des médias vis-à-vis des annonceurs, des partis politiques ou d'autres lobbys est primordiale. À cet égard, le contrôle accru sur la recommandation algorithmique permise par la fabrication d'exemples adversariaux est préoccupant. Dans une vision plus participative de la démocratie, le rôle des médias dépasse le simple devoir d'information et s'oriente davantage vers une éducation active de citoyens engagés (Helberger, 2019). De ce fait, des systèmes de recommandation « démocratiques » se doivent d'opérer une sélection honnête et représentative de l'information. Il s'agit ainsi d'éviter les bulles de filtre, les chambres d'échos et la polarisation des débats, écueils qui tous peuvent être amplifiés par des exemples adversariaux. À cet égard, nous renvoyons aux travaux de Perra et Rocha (2019) qui étudient la dynamique des opinions sur les réseaux sociaux et comment une légère inflexion peut suffire à les influencer.

Dans le cadre de notre deuxième étude de cas, à savoir l'évaluation automatique de la fiabilité d'une information, l'impact qu'auraient des comportements adversariaux est également manifeste et s'ajoute à la difficulté –déjà évoquée plus haut– inhérente à la définition même de la notion de fiabilité d'une information (Tandoc, Lim et al., 2018). D'autres exemples du même type existent, qui combinent également, d'une part, une nécessaire automatisation des processus face à l'augmentation drastique du volume d'informations à traiter ; et, d'autre part, une complexité (voire une impossibilité) à définir des critères objectifs et unanimes pour piloter cette automatisation. Nous pouvons par exemple citer la modération des commentaires en ligne dont la quantité croissante requiert une part d'automatisation (Arnt et Zilberstein, 2003). Là aussi, Binns, Veal et al. (2017) montrent comment cette tâche est rendue ardue par les différentes conceptions du caractère offensant d'un commentaire. Plus récemment, des outils d'analyse automatique de l'opinion et des marqueurs de subjectivité d'un article ont été développés (Carlebach, Cheruvu et al., 2020) dans la perspective d'une meilleure contextualisation de l'information et de la lutte contre la polarisation des débats en ligne. Dans chacun de ces exemples, il est aisé d'identifier l'intérêt que pourrait avoir un individu ou un groupe d'individus à tromper le système automatisé en question, qu'il s'agisse de faire passer pour fiable une information qui ne l'est pas, faire publier un commentaire offensant, ou diffuser un article d'opinion dans une rubrique supposée factuelle.

De manière générale, nous pouvons donc constater que les comportements adversariaux (et plus globalement le problème de la sécurité des algorithmes d'apprentissage automatique) nécessitent l'extension des critères d'évaluation de la performance d'un algorithme. Ainsi, le comportement attendu des algorithmes doit être garanti également dans des contextes adversariaux, c'est-à-dire des contextes dans lesquels les données à traiter sont contrôlées par un adversaire. Insistons sur le fait que cette exigence de robustesse est complémentaire à celle de transparence, qui reste un élément nécessaire à la bonne compréhension des objectifs des personnes ayant conçu l'algorithme. Il s'agit davantage d'une exigence technique supplémentaire, à prendre en compte dans la conception d'un algorithme. Et la fabrication d'exemples adversariaux apparaît par conséquent comme un outil légitime d'investigation et de validation de cette responsabilité algorithmique (*algorithmic accountability reporting*, tel que défini par Diakopoulos, 2019).

Défis techniques et contre-mesures

D'un point de vue technique, les résultats présentés dans cet article ouvrent plusieurs pistes de recherche. Par exemple, la génération automatique d'exemples adversariaux qui soient difficiles à repérer est un problème intéressant. Notons à cet égard que le mimétisme de la classe cible n'est pas toujours un objectif : ce n'était pas le cas pour le système de recommandation d'articles (l'objectif n'était pas qu'un article classé Culture, pour passer à la classe Sport, utilise ostensiblement du vocabulaire sportif), mais c'était cependant le cas pour le détecteur de désinformation (l'objectif était bien là de faire apparaître comme fiable un article non fiable). Dans tous les cas, en revanche, l'idée est de ne pas altérer le contenu sémantique de l'article original. Comprendre dès lors comment automatiser la fabrication de tels exemples sans altérer ce contenu sémantique permettrait d'appréhender la façon dont les exemples adversariaux pourraient être généralisés et déployés à grande échelle. Notons qu'une telle génération automatique bénéficierait d'une transparence de l'algorithme visé : connaître les détails de l'algorithme d'apprentissage et les différents paramètres du modèle sous-jacent faciliterait la tâche de fabrication (mais aussi de détection) des exemples adversariaux.

Un autre défi est d'analyser la difficulté avec laquelle ces exemples adversariaux sont fabriqués en fonction de la taille du corpus de données utilisé pour l'entraînement. La raison pour laquelle un mot aussi neutre que « normalement » était suffisant pour faire basculer un article dans notre système de classification automatique était liée à une variation significative des occurrences de ce mot entre les différentes classes. On peut supposer qu'une augmentation du nombre d'articles utilisés pour entraîner le système tendrait à équilibrer les probabilités d'occurrence de ces mots neutres. Cela étant, le spectre des critères utilisés par un algorithme d'apprentissage automatique étant virtuellement infini, il est tout à fait envisageable que d'autres motifs discriminants, mais apparemment neutres, soient perçus par l'algorithme et puissent être ensuite utilisés dans des exemples adversariaux.

Enfin, nous pouvons analyser les contre-mesures techniques à opposer aux comportements adversariaux. Rappelons à cet égard que nous n'affirmons pas que les exemples adversariaux constituent directement une menace critique pour n'importe quelle application d'apprentissage automatique. Nous souhaitons uniquement mettre en évidence le fait que ce risque devrait être pris en compte dans l'évaluation des performances d'un algorithme, ce qui soulève la question des contre-mesures possibles. À ce stade de la recherche dans ce domaine, un certain nombre d'heuristiques peut être envisagé.

Une approche assez intuitive est de combiner différents modèles (comme NB, MLP ou RNN dans notre cas), en postulant que les exemples adversariaux sont spécifiques à un modèle donné et que, par conséquent, une combinaison de plusieurs d'entre eux évitera des classifications erronées. Cependant, plusieurs expériences montrent que les exemples adversariaux ont tendance à se transférer entre les modèles, ce qui est par ailleurs une conséquence attendue de la bonne généralisation de ces modèles (Tramèr, Papernot et al., 2017). Une approche plus prometteuse semble être « *l'entraînement adversarial* » (Tramèr, Kurakin et al., 2018) qui consiste à injecter des exemples adversariaux durant la phase d'entraînement du système afin d'accroître sa robustesse.

Pour autant, nous sommes encore loin d'une situation où des garanties techniques fortes de cette robustesse peuvent être fournies et où les risques posés par des exemples adversariaux pourraient être ignorés. L'étude de ces contre-mesures dans le domaine du journalisme computationnel en particulier est une question de recherche intéressante qui reste encore à explorer.

Extension du rôle des journalistes

Comme l'a observé Coddington, l'utilisation du raisonnement computationnel ou de l'abstraction des données pour la réalisation de tâches complexes dans les processus de production d'information ne semble pas avoir d'équivalent dans le journalisme précédant l'ère informatique (2015, 344). Par conséquent, et malgré la tendance croissante à établir des liens entre les pratiques des rédactions et les connaissances en informatique et en ingénierie, certaines problématiques techniques issues de ces domaines restent encore largement méconnues. Le cas des exemples adversariaux présenté dans cet article en est une illustration et met en lumière à quel point l'automatisation et l'apprentissage automatique soulèvent des questions que les journalistes et les chercheurs en sciences sociales n'ont pas l'habitude de considérer. Il ne s'agit pas seulement d'une conséquence du caractère émergent de ces techniques : cette situation est également due au fait que les personnes impliquées dans le domaine journalistique ne pensent pas nécessairement leur travail en termes computationnels. Ainsi, il est plus courant d'étudier comment les technologies peuvent menacer ou faciliter des pratiques journalistiques établies, ou une certaine mission d'information, plutôt que de mobiliser un autre point de vue ou une autre expertise permettant de mettre en lumière des problématiques jusque-là insoupçonnées.

Ce travail confirme donc à quel point la compréhension du fonctionnement des algorithmes et de la manière dont ils modifient les frontières du journalisme (Carlson et Lewis, 2015) pourrait tirer profit d'une plus grande intégration de différents champs de recherche. Comme le suggère Coddington (2015), cette intégration ne devrait pas se cantonner aux dimensions techniques et matérielles, car celles-ci sont encadrées par les valeurs et orientations épistémologiques que toute discipline développe. Ainsi, la méthodologie présentée ici consistant à contester et affiner une définition (celle de responsabilité algorithmique) à l'aide de contre-exemples, tout à fait classique en cryptographie, est un exemple de ce que ce type d'intégration peut apporter.

Le journalisme, qui est fréquemment présenté comme un domaine au périmètre flou et fluctuant, sensible aux influences des autres, a depuis longtemps démontré une capacité à intégrer des éléments de diverses cultures professionnelles (voir par exemple Lewis et Usher, 2013, pour ce qui concerne l'automatisation). La complexité technique et les enjeux politiques inhérents au processus algorithmique rendent cette capacité d'ouverture particulièrement souhaitable au regard des différentes épistémologies (Ward, 2015, 2018) et des rôles normatifs importants joués par les journalistes (chercher, vérifier et diffuser de l'information).

De manière plus générale, les exemples adversariaux, et le caractère limité des contre-mesures techniques existantes à ce jour, font naturellement écho à la société du risque théorisée par Beck (1992), c'est-à-dire « *une société où nous vivons de plus en plus sur une frontière de haute technologie que personne ne comprend complètement* ». Dans ce type de société, chaque organisation s'emploie à gérer des risques, en lien avec un certain nombre

de valeurs que l'organisation souhaite défendre. Dans le contexte du journalisme computationnel, le niveau de risques qui pourra être toléré dépendra donc des valeurs qui se trouveront menacées par les biais induits par les adversaires. Notons à cet égard que si des risques physiques provoquent des dommages bien visibles, les risques posés par les comportements adversariaux sont davantage pernicieux dans le sens où les biais qu'ils induisent (et les dommages qui en résultent) peuvent passer totalement inaperçus. Nous renvoyons sur ce point aux travaux de Rouvroy et Berns sur la gouvernementalité algorithmique (2013) : les algorithmes, par le fait qu'ils façonnent notre environnement le plus souvent à notre insu et sans nous donner à voir les alternatives possibles, sont une forme de gouvernement « en creux », une norme qui ne dit pas son nom et qui avance masquée. Ainsi, l'invisibilisation des dommages provoqués par des comportements adversariaux souligne d'autant plus l'enjeu majeur que représente une responsabilité algorithmique robuste.

Dans ce contexte, il est intéressant de s'arrêter un instant sur le rôle que peuvent jouer les journalistes dans la prévention des risques liés à l'automatisation et aux comportements adversariaux en particulier. En effet, une interaction directe entre la technologie et des usagers finaux non informés mène à des risques qu'il est difficile de prévenir et imposerait donc a priori l'application d'un principe de précaution. Mais la situation des médias d'information est différente dans le sens où les journalistes peuvent jouer un rôle d'interface entre la production automatisée d'informations et les publics. Ainsi, il serait souhaitable d'étendre leur rôle traditionnel de garant de l'information à celui de garant des décisions algorithmiques, en mesure de tester la qualité des données collectées et la robustesse des algorithmes qui les manipulent. Ce rôle d'interface permettrait une clarification des questions de responsabilité : le/la journaliste demeure auteur/trice et responsable de l'information diffusée, et il/elle peut en rendre compte.

Conclusion

Le travail présenté ici a exploré la question de la responsabilité algorithmique dans le domaine journalistique, et les garanties qu'il est possible d'obtenir sur la pertinence des décisions automatiques prises dans les processus de production de l'information. Les deux études de cas que nous avons présentées nous ont permis de mettre en lumière les limites techniques de l'idéal de transparence souvent présentée comme une garantie suffisante de responsabilité algorithmique. Nous avons ainsi montré que même si l'analyse des données d'entraînement et des paramètres de l'algorithme indique qu'il produira, dans des conditions normales d'utilisation, des résultats conformes aux attentes, cela ne suffit pas pour avoir la garantie qu'il se comportera toujours de cette manière. Il est en effet possible de fabriquer des exemples adversariaux qui lui fassent prendre des décisions qui ne soient pas conformes à l'intention des personnes l'ayant conçu.

Sur la base de nos résultats et de l'analyse de leurs conséquences concrètes dans le domaine du journalisme computationnel, nous suggérons que la performance d'un algorithme puisse également inclure une évaluation de sa robustesse face à de tels comportements adversariaux et, plus généralement, que soit prise en compte la sécurité de l'apprentissage automatique dans la conception des processus automatisés de production de l'information. Pour autant, cette robustesse n'est pas aisée à garantir et nous avons mis en lumière les limitations des contre-mesures techniques existant à ce stade.

Dans ce contexte, si notre travail encourage une conception algorithmique aussi robuste que possible, il plaide également pour une extension du rôle des journalistes et des membres de rédaction comme garants de la pertinence des décisions automatiques. Cette implication accrue est parfaitement alignée avec la perspective de « *garder l'humain dans la boucle* », mise en avant par Milosavljević et Vobič (2019), et qui souligne que malgré l'importance grandissante de l'automatisation, les journalistes restent les agents principaux du processus de production d'information et de sa réinvention continuelle. Elle est également en phase avec l'analyse de Bucher qui rappelle que si les algorithmes transforment le journalisme, ils n'éliminent pas la nécessité du jugement humain et de l'expertise dans le travail des médias d'information. « *Les algorithmes déplacent, redistribuent, et façonnent les nouvelles manières d'être un travailleur de l'information* » (Bucher, 2018, 145). De nombreux exemples existent mettant en lumière cette hybridation progressive des pratiques et reposant encore largement sur la pertinence du jugement humain : voir par exemple les travaux de Park, Sachar et al. (2016) sur la modération semi-automatique des commentaires en ligne.

Cette extension du rôle des journalistes conforte une conception de la responsabilité algorithmique davantage centrée sur la personne utilisant un système automatisé. Cette conception apparaît comme complémentaire à une approche ciblant la conception technique uniquement, et permettrait de combler ses lacunes. Concrètement, une approche pragmatique de la responsabilité algorithmique consisterait donc à concevoir des processus algorithmiques qui soient les plus transparents et robustes possibles, tout en restant conscient de ses limitations, et à se reposer ensuite sur une conception de la responsabilité centrée sur l'usager. Cette approche permettrait un usage à la fois ambitieux et contrôlé des nouvelles technologies. ■

Antonin Descampe est chargé de cours à l'Université catholique de Louvain et membre de l'Observatoire de recherche sur les médias et le journalisme, François-Xavier Standaert est professeur au groupe Crypto de l'Université catholique de Louvain et maître de recherche FNRS.

Références

Ananny, Mike et Crawford, Kate (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989.

Anderson, CW (2013). Towards a sociology of computational and algorithmic journalism. *New Media & Society*, 15(7), 1005-1021.

Arnt, A. et Zilberstein, S. (2003). Learning to perform moderation in online forums. Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003).

Barocas, Solon, Hardt, Moritz et Narayanan, Arvind (2017). Fairness and machine learning. *Nips tutorial*, 1, 2017.

Barreno, Marco, Nelson, Blaine, Joseph, Anthony D. et Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, 81(2), 121-148.

Beck, Ulrich, Lash, Scott et Wynne, Brian (1992). *Risk society: Towards a new modernity*. Sage.

Biggio, Battista, Nelson, Blaine et Laskov, Pavel (2012). Poisoning attacks against support vector machines. arXiv.org.

Binns, Reuben, Veale, Michael, Van Kleek, Max et Shadbolt, Nigel (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. Cham.

Bodó, Balázs (2019). Selling news to audiences – A qualitative inquiry into the emerging logics of algorithmic news personalization in European quality news media. *Digital Journalism*, 7(8), 1054-1075.

Bogaert, Jérémie, Carbonnelle, Quentin, Descampe, Antonin et Standaert, François-Xavier (2021). Can fake news detection be accountable? The adversarial examples challenge. 41st WIC Symposium on Information Theory in the Benelux.

Broussard, Meredith (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Broussard, Meredith, Diakopoulos, Nicholas, Guzman, Andrea L., Abebe, Rediet, Dupagne, Michel et Chuan, Ching-Hua (2019). Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, 96(3), 673-695.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, et al. (2020). *Language Models are Few-Shot Learners*. En ligne : arxiv.org/abs/2005.14165.

Bucher, Taina (2018). *If...then: Algorithmic power and politics*. Oxford University Press.

Burrell, Jenna (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).

Carlebach, Mark, Cheruvu, Ria, Walker, Brandon, Magalhaes, Cesar Ilharco et Jaume, Sylvain (2020). News aggregation with diverse viewpoint identification using neural embeddings and semantic understanding models. *Proceedings of the 7th Workshop on Argument Mining*, 59-66.

Carlson, Matt et Lewis, Seth C. (2015). *Boundaries of journalism: Professionalism, practices and participation*. Routledge.

Coddington, Mark (2015). Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital journalism*, 3(3), 331-348.

Crain, Matthew (2018). The limits of transparency: Data brokers and commodification. *New Media & Society*, 20(1), 88-104.

Crawford, Kate et Schultz, Jason (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*, 55(1), 93-128.

Dagiral, Éric et Parasie, Sylvain (2017). La « science des données » à la conquête des mondes sociaux : ce que le « Big Data » doit aux épistémologies locales. *Big data et*

traçabilité numérique. Les sciences sociales face à la quantification massive des individus, Paris, Collège de France, 85-104.

Datta, Amit, Tschantz, Michael Carl et Datta, Anupam (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. arXiv.org

Descampe, Antonin, Massart, Clément, Poelman, Simon, Standaert, François-Xavier et Standaert, Olivier (2021). Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. *AI & Society* [à paraître].

Diakopoulos, Nicholas (2015). Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398-415.

Diakopoulos, Nicholas (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

Eykholt, Kevin, Evtimov, Ivan, Fernandes, Earlence, Li, Bo, Rahmati, Amir, Xiao, Chaowei, Prakash, Atui, Tadayoshi, Kohno, Song, Dawn (2018). Robust physical-world attacks on deep learning visual classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1625-1634.

Goodfellow, Ian J., Shlens, Jonathon et Szegedy, Christian (2015). Explaining and harnessing adversarial examples. arXiv.org.

Goodfellow, Ian, McDaniel, Patrick et Papernot, Nicolas (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7), 56-66.

Helberger, Natali (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), 993-1012.

Hochreiter, Sepp et Schmidhuber, Jürgen (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Karppinen, Kari (2018). Journalism, pluralism, and diversity. Dans Tim P. Vos (dir.), *Journalism,* 493-510. De Gruyter.

Katz, Jonathan et Lindell, Yehuda (2020). *Introduction to modern cryptography*. CRC press.

Kormelink, Tim Groot et Meijer, Irene Costera (2014). Tailor-made news. *Journalism Studies*, 15(5), 632-641.

Kroll, Joshua A., Huey, Joanna, Barocas, Solon, Felten, Edward W., Reidenberg, Joel R., Robinson, David G. et Yu, Harlan (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705.

Kunert, Jessica et Thurman, Neil (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010-2016. *Journalism Practice*, 13(7), 759-780.

Lazer, David M. J., Baum, Matthew A., Benkler, Yochai, Berinsky, Adam J., Greenhill, Kelly M., Menczer, Filippo, Metzger, Miriam J., Nyhan, Brendan, Pennycook, Gordon, Rothschild, David, Schudson, Michael, Sloman, Steven A., Sunstein, Cass R., Thorson, Emily A., Watts, Duncan J. et Zittrain, Jonathan L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.

Lepri, Bruno, Oliver, Nuria, Letouzé, Emmanuel, Pentland, Alex et Vinck, Patrick (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611-627.

Lewis, SC et Westlund, O. (2016). Mapping the human–machine divide in journalism. *The SAGE Handbook of Digital Journalism*, 341-353.

Lewis, Seth C., Guzman, Andrea L. et Schmidt, Thomas R. (2019). Automation, journalism, and human–machine communication: rethinking roles and relationships of humans and machines in news. *Digital Journalism*, 7(4), 409-427.

Lewis, Seth C. et Usher, Nikki (2013). Open source and journalism: toward new frameworks for imagining news innovation. *Media, Culture & Society*, 35(5), 602-619.

Liang, Bin, Li, Hongcheng, Su, Miaoqiang, Bian, Pan, Li, Xirong et Shi, Wenchang (2018). Deep text classification can be fooled. Stockholm, Sweden. *Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}*.

Mikolov, Tomas, Chen, Kai, Corrado, Greg et Dean, Jeffrey (2013). Efficient estimation of word representations in vector space. arXiv.org.

Milano, Silvia, Taddeo, Mariarosaria et Floridi, Luciano (2020). Recommender systems and their ethical challenges. *AI & Society*, 35(4), 957-967.

Milosavljević, Marko et Vobič, Igor (2019). Human still in the loop. *Digital Journalism*, 7(8), 1098-1116.

Morris, John, Lifland, Eli, Lanchantin, Jack, Ji, Yangfeng et Qi, Yanjun (2020). Reevaluating adversarial examples in natural language. arXiv.org.

Nielsen, Rasmus Kleis (2016). The many crises of Western journalism: A comparative analysis of economic crises, professional crises, and crises of confidence. *The crisis of journalism reconsidered*, 77-97.

Park, Deokgun, Sachar, Simranjit, Diakopoulos, Nicholas et Elmqvist, Niklas (2016). Supporting comment moderators in identifying high quality online news comments. Dans *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1114-1125.

Perra, Nicola et Rocha, Luis E. C. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9(1), 7261.

Ribeiro, Marco Tulio, Singh, Sameer et Guestrin, Carlos (2018). Semantically equivalent adversarial rules for debugging NLP models. Melbourne, Australia. *ACL 2018*.

Rouvroy, Antoinette et Berns, Thomas (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Reseaux*, 177(1), 163-196.

Rumelhart, David E., Hinton, Geoffrey E. et Williams, Ronald J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

Sandvig, Christian, Hamilton, Kevin, Karahalios, Karrie et Langbort, Cedric (2014). Auditing algorithms: Research methods for detecting discrimination on internet

platforms. *Data and discrimination: Converting critical concerns into productive inquiry*, 22, 4349-4357.

Sato, Motoki, Suzuki, Jun, Shindo, Hiroyuki et Matsumoto, Yuji (2018). Interpretable adversarial perturbation in input embedding space for text. Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18). 4323-4330.

Schuster, Mike et Paliwal, Kuldip K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673-2681.

Spärck Jones, Karen (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.

Steinhardt, Jacob, Koh, Pang Wei et Liang, Percy (2017). Certified defenses for data poisoning attacks. arXiv.org.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian et Fergus, Rob (2014). Intriguing properties of neural networks. arXiv.org.

Tandoc, Edson C., Lim, Zheng Wei et Ling, Richard (2018). Defining "fake news". *Digital Journalism*, 6(2), 137-153.

Thurman, Neil, Lewis, Seth C. et Kunert, Jessica (2019). Algorithms, automation, and news. *Digital Journalism*, 7(8), 980-992.

Thurman, Neil, Moeller, Judith, Helberger, Natali et Trilling, Damian (2019). My friends, editors, algorithms, and I. *Digital Journalism*, 7(4), 447-469.

Tramèr, Florian, Kurakin, Alexey, Papernot, Nicolas, Goodfellow, Ian, Boneh, Dan et McDaniel, Patrick (2020). Ensemble adversarial training: Attacks and defenses. arXiv.org.

Tramèr, Florian, Papernot, Nicolas, Goodfellow, Ian, Boneh, Dan et McDaniel, Patrick (2017). The space of transferable adversarial examples. arXiv.org.

Vargo, Chris J., Guo, Lei et Amazeen, Michelle A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20(5), 2028-2049.

Wang, William Yang (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv.org.

Ward, Stephen J. A. (2015). *Radical media ethics: A global approach*. John Wiley & Sons.

Ward, Stephen J. A. (2018). Epistemologies of journalism. Dans Tim P. Vos (dir.), *Journalism* (p. 63-82). De Gruyter.

Wing, Jeannette M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1881), 3717-3725.

Xue, Mingfu, Yuan, Chengxiang, Wu, Heyi, Zhang, Yushu et Liu, Weiqiang (2020). Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, 8, 74720-74742.

Zellers, Rowan, Holtzman, Ari, Rashkin, Hannah, Bisk, Yonatan, Farhadi, Ali, Roesner, Franziska et Choi, Yejin (2019). Defending against neural fake news. *Neurips*, 9051-9062.

Zhou, Xinyi et Zafarani, Reza (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM computing surveys*, 53(5), 109:1-109:40.