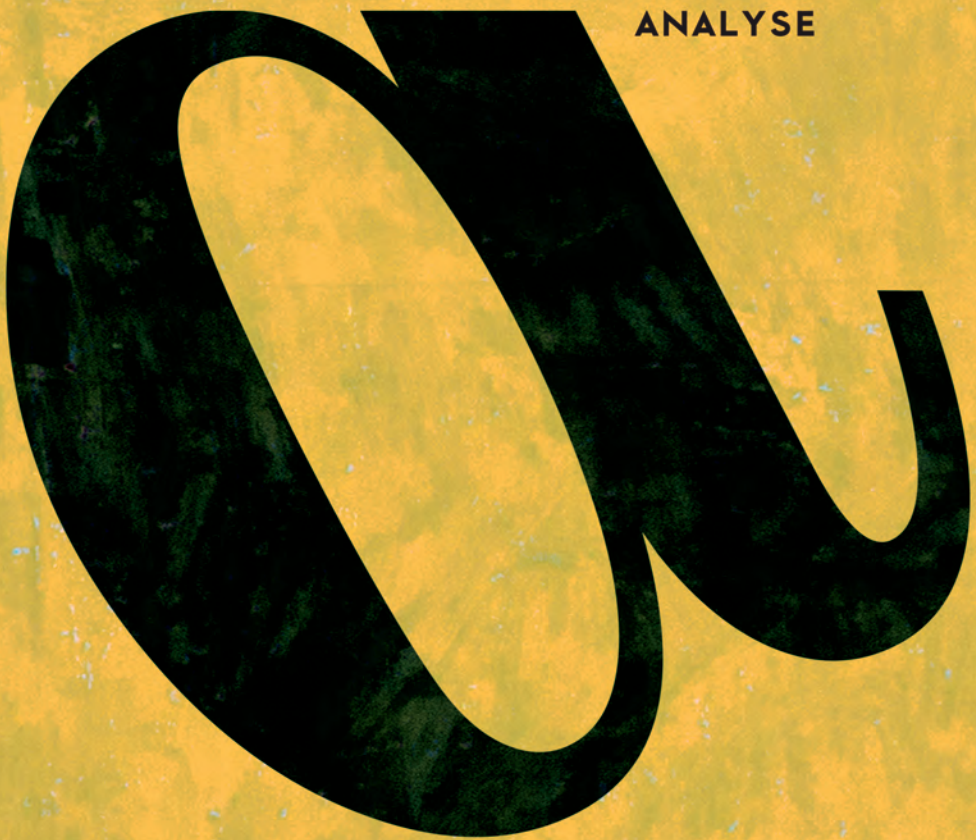
POUR UNE EXPLICITATION GÉNÉRALISÉE DU DÉVELOPPEMENT ALGORITHMIQUE : CONCEPTION-PROBLÉMATISATION ET DIFFUSION-PERFORMANCE

FLORIAN **JATON**



Comment reprendre collectivement la main sur les algorithmes qui peuplent nos quotidiens ? Dans cette note, Florian Jaton rend compte d'une série d'outils théoriques – issus pour la plupart des Science and Technology Studies (STS) – soutenant l'inspection plus avant des algorithmes au moment de leur conception ainsi qu'au moment de leur diffusion. Ces deux lignes de réflexion – nommées ici "conception-problématisation" et "diffusion-performance" – constituent autant d'accroches pour une meilleure préhension des dispositifs algorithmiques, une condition nécessaire à leur encadrement.

Société algorithmique ! Si l'expression a été justement critiquée – il est en effet difficile d'imaginer une société réduite aux seules relations médiées par des dispositifs numériques[1] – elle a néanmoins le mérite de rappeler l'omniprésence des méthodes informatiques de calcul et leur extension semble-t-il irrésistible[2]. Réseaux sociaux, logiciels de bureautique, générateurs d'images, systèmes de surveillance, correcteurs automatiques : pas un seul jour ne passe sans que nos cours d'action ne se mêlent à des centaines d'algorithmes inscrits dans les profondeurs de nos ordinateurs.

Depuis au moins une dizaine d'années, cet état de fait nourrit des volontés de régulation, notamment aux États-Unis, au Royaume-Uni, en Chine et en Europe[3]. L'argumentaire est limpide et légitime : en tant que moteur de ce qu'il est convenu d'appeler – un peu abstraitement – *le numérique* (et récemment plus abstraitement encore, l'*intelligence artificielle*), les algorithmes doivent respecter les cadres légaux des sociétés qu'ils irriguent. Mais la chose est loin d'être facile : entre les puissants lobbys industriels qui pèsent de tout leur poids sur les commissions parlementaires – en mobilisant notamment l'épouvantail de la perte de compétitivité au sein d'un secteur technologique à haute valorisation – et le caractère abstrait de la notion d'algorithme – difficile en effet de *voir concrètement* ce à quoi cette notion réfère – les efforts de régulation traversent bien des vicissitudes. D'où l'importance de poursuivre les réflexions quant aux moyens à mettre en place pour reprendre effectivement la main sur les algorithmes, ces boîtes noires qui nourrissent les décisions, les amours et les peines.

Afin de contribuer à cette entreprise collective d'encadrement, je souhaiterais ici présenter une série de travaux – souvent rassemblés sous la bannière des *Science & Technology Studies* (STS)[4] – qui ont souligné le besoin d'inspecter plus avant les algorithmes au moment de leur conception

1 Bogost Ian, « The Cathedral of Computation », *The Atlantic*, 15 janvier 2015, URL : http://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/, consulté le 29 Avril 2016.

2 Mazzotti Massimo, « Algorithmic Life », *Los Angeles Review of Books*, 22 janvier 2017, URL : https://lareviewofbooks.org/article/algorithmic-life/, consulté le 10 mai 2019.

3 Pour un résumé des tentatives de régulation en cours, notamment en Europe, aux États-Unis, au Royaume-Uni et en Chine, voir le récent article de Malingre Virginie, Ducourtieux Cécile et Smolar Piotr, « Intelligence artificielle : la course à la régulation entre grandes puissances », *Le Monde*, 9 juin 2023. URL : https://www.lemonde.fr/international/article/2023/06/09/intelligence-artificielle-la-course-a-la-regulation-entre-grandes-puissances_6176823_3210.html, consulté le 10 juin 2023.

4 Comme le disent Peter Dear et Sheila Jasanoff, ce qui relie les chercheur·euse·s de cette communauté hétérogène des *Science & Technology Studies* (STS) est la conviction que « depuis les travaux de Ludwik Fleck, Thomas Kuhn, et David Bloor, la science n'est pas le froid royaume de l'empirisme logique et des principes premiers, que nous préexistons nos connaissances du monde (tout comme le monde préexiste nos connaissances), que la matérialité est centrale à la formation et à la mise à l'épreuve des vérités scientifiques, et que les sciences et les dynamiques des pratiques scientifiques et techniques sont un terrain fertile à l'analyse sociale, politique et éthique. » Dear Peter et Jasanoff Sheila, « Dismantling Boundaries in Science and Technology Studies », *Isis*, t. CI, n°4, 2010, p. 761 – ma traduction.

ainsi qu'au moment de leur diffusion. La première ligne axée « conception » s'attache à rendre explicite les postulats qui sous-tendent la formation de problèmes à même d'être résolus computationnellement. La deuxième ligne axée « diffusion » s'attache à rendre explicite les caractéristiques des performances des algorithmes exploitables. L'un dans l'autre, ces deux lignes – que je nommerai ici « conception-problématisation » et « diffusion-performance » – constituent autant d'accroches pour une meilleure préhension des dispositifs algorithmiques, une condition nécessaire à leur encadrement.

CONCEPTION–PROBLÉMATISATION

Les algorithmes – entendus ici comme méthodes informatiques de calcul – ne tombent pas du ciel (même si on en a parfois l'impression) : ce sont des constructions situées, humaines et culturelles[5]. Et parmi les nombreuses pratiques et matériaux dont ces constructions ont besoin pour advenir, il y a des bases de données numériques, souvent appelées *ground truth* dans la littérature spécialisée[6]. En mettant en relation des données d'entrée (ce que l'algorithme-en-devenir devra traiter) et des cibles de sortie (ce que l'algorithme-en-devenir devra reproduire), ces bases de données *ground truth* constituent le substrat fondamental des algorithmes : elles définissent matériellement le *problème* qu'un algorithme devra résoudre (voir figure 1).

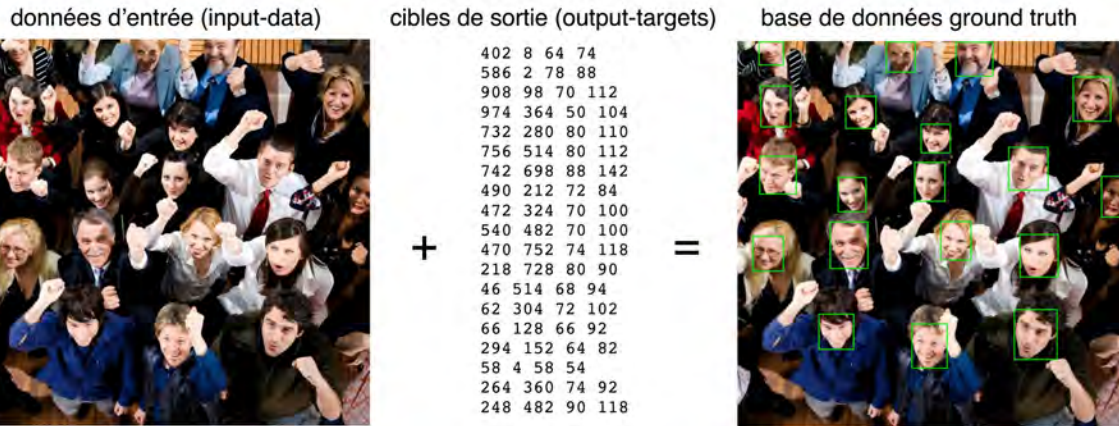


Figure 1. Tiré de Yang Shuo, Luo Ping, Loy Chen Change et Tang Xiaoou, « WIDER FACE: A face detection benchmark », in *2016 IEEE conference on computer vision and pattern recognition*, New York, IEEE Press, 2016. ; échantillon de la base de données *ground truth* WIDER FACE pour la recherche en détection faciale. À gauche, une parmi

5 Seaver Nick, « Algorithms as culture: Some tactics for the ethnography of algorithmic systems », *Big Data & Society*, t. IV, n°2, 2017.

6 Jaton Florian, « We get the algorithms of our ground truths: Designing referential databases in digital image processing », *Social Studies of Science*, t. XLVII, n°6, 2017.

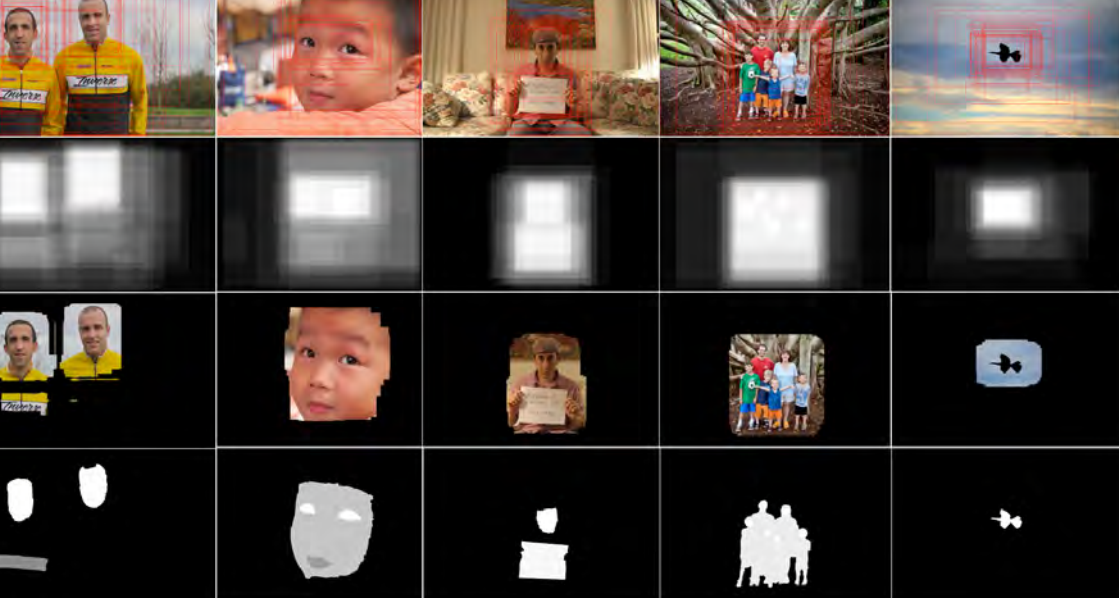
les 32 203 images de ce jeu de données accessible publiquement. Au milieu, les annotations de visage pour cette image spécifique, produites par des humains (en l'occurrence, les signataires de l'article académique introduisant la *ground truth*). Comme chaque annotation appartient à l'espace de coordonnées de l'image, elle peut être exprimée par un ensemble de quatre valeurs, les deux premières exprimant la position de départ de l'annotation le long des axes x et y, la troisième exprimant le nombre de pixels de largeur, la quatrième exprimant le nombre de pixels de hauteur. Ces informations, qui correspondent aux rectangles verts de l'image de droite, ont été produites manuellement par un annotateur humain et vérifiées par deux autres (Yang et al., 2016 : 5527). En tant que telles, ces annotations constituent les cibles de sorties (*output-targets*) de l'image pour ce qui est de la détection faciale ; elles viennent s'additionner aux données d'entrée afin de fournir quelque chose à apprendre et à formuler. Les données d'entrée (*input-data*) et leurs cibles de sorties (*output-targets*) peuvent ensuite être utilisées pour construire des algorithmes de détection faciale, la base de données *ground truth* opérant comme la liste des meilleures réponses pour cette tâche précise. Source: Jaton Florian, « Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application », *Big Data & Society*, t. VIII, n°1, 2021, p. 3.

Pour autant, tout comme les algorithmes qu'elles participent à engendrer, les bases de données *ground truth* – et donc les problèmes qu'elles définissent – ne préexistent pas : elles émanent de processus de construction concrets et situés qui impliquent plus ou moins de personnes, d'efforts et de ressources (voir figure 2). Rarement mis en avant dans les discours promotionnels des géants de la tech, ces processus arbitraires (car ils sont menés par des individus situés socialement) et contingents (car ils pourraient être menés différemment) dits de *ground-truthing* sont pourtant centraux : sans eux, pas de bases de données *ground truth*, et donc *in fine*, pas d'algorithmes (car impossible dès lors de les construire[7]).

Figure 2 (page suivante). Exemple de quelques étapes d'une construction d'une base de données *ground truth*. Il s'agit ici de « reconnaître la saillance », ce qui équivaut à segmenter les contours des éléments (visages, objets, animal) qui attirent le regard du plus grand nombre de personnes au sein d'une photographie. Les différentes opérations impliquées aboutissent à la fabrication d'un certain cadrage du problème de la détection de saillance. D'autres problématisations auraient pu advenir. Les algorithmes qui dériveront de cette base de données porteront en eux – et dès lors promouvront – cette problématisation spécifique – et arbitraire – de la notion de saillance. Source : Jaton Florian, *The Constitution of Algorithms*, op. cit., p. 74.

Les activités de *ground-truthing* irriguent fortement nos sociétés de plus en plus informatisées : comme ces activités définissent les problèmes que les algorithmes-en-devenir vont être amenés à résoudre, elles sont les

7 Jaton Florian, *The constitution of algorithms : Ground-truthing, programming, formulating*, Cambridge, Mass., MIT Press, 2021.



conditions d'existence de ce qui anime les dispositifs numériques avec lesquels nous ne cessons d'interagir (volontairement ou non). Comment dès lors rendre justice à ces activités et, le cas échéant, souligner la responsabilité qu'incombe à celles et ceux qui s'y adonnent ? C'est dans ce but qu'Edward Kang a récemment proposé la notion de *ground-truth tracing*[8], soit le compte rendu minutieux, et potentiellement systématique, des moments durant lesquels des individus travaillant dans des laboratoires d'informatique académiques ou industriels s'engagent dans un processus de *ground-truthing*.

Bien qu'encore spéculatif en l'état, le *ground-truth tracing* opérerait comme suit. Il s'agirait dans un premier temps de documenter – idéalement sur le vif – les dynamiques à l'œuvre durant la construction d'une base de données *ground truth*. Ensuite, dans un deuxième temps, il s'agirait de s'appuyer sur ce compte rendu textuel pour *qualifier* le travail de problématisation tel qu'il a été effectué. Pour ce faire, Kang propose de mobiliser deux critères. Le premier a trait au degré d'univocité du problème tel qu'exprimé dans la *ground truth*, qui renvoie par exemple à l'existence ou non de normes et de standards partagés. Grande est en effet la différence entre construire une *ground truth* pour soutenir la construction d'algorithmes capables de détecter des défauts géométriques en usinage de pièces métalliques et construire une *ground truth* pour soutenir la construction d'algorithmes capables de détecter des émotions à partir d'expressions vocales. La première situation est en effet relativement *univoque* et calibrée par des normes et des critères de qualité ; la deuxième est une situation for-

8 Kang Edward, «Ground truth tracings (GTT): On the epistemic limits of machine learning», *Big Data & Society*, t. X, n°1, 2023.

tement *équivoque*, changeante et dépendante de multiples facteurs sociaux tout à fait contingents. Si chacune des deux situations, lorsque traduite en base de données *ground truth*, pourrait *donner l'impression* d'une univocité non-problématique, seule la première l'est réellement, chose que seul un compte rendu détaillé de *ground-truth tracing* serait capable de rapporter et de signaler précisément.

Le deuxième critère proposé par Edward Kang pour qualifier le travail de problématisation tel que rapporté par une entreprise de *ground-truth tracing* a trait au degré d'enjeu social. En effet, pour reprendre les exemples proposés plus haut, si les enjeux liés à la construction d'une *ground truth* pour fonder la détection de défauts de pièces métalliques sont certainement importants pour l'entreprise concernée (voire pour le secteur auquel elle appartient), ces enjeux restent plutôt maigres pour le reste de la société qui ne se verrait dès lors pas fortement impactée. Pour ce qui est de la *ground truth* pour aider à la détection d'émotions à partir d'expressions vocales, la situation est différente : une fois des algorithmes conçus à partir de cette *ground truth*, ceux-ci pourraient être mobilisés dans des secteurs très différents, allant du recrutement d'employés (par exemple pour « repérer » des traits de caractères particuliers) à l'enquête policière (par exemple pour « identifier » des témoignages mensongers), en en passant par le diagnostic psychiatrique (par exemple pour « déceler » des états pathologiques[9]). En ce sens, l'enjeu social sous-jacent à la définition de cette *ground truth* pour la détection d'émotions à partir d'expressions vocales est a priori plus important que l'enjeu social sous-jacent à la *ground truth* pour la détection de défauts de pièces métalliques.

L'un dans l'autre, ces deux critères permettraient de positionner les narratifs de *ground-truth tracing* au sein d'un système de coordonnées (dont la gradation reste à imaginer) permettant de visualiser les processus de problématisation selon leurs différents degrés d'équivocité et d'enjeu social (voir figure 3). Ce type de représentation – pour l'heure tout à fait spéculative – pourrait en retour permettre d'identifier des problématisations algorithmiques plus ou moins risquées, selon leur position au sein de ce tableau. Une problématisation placée proche de l'origine serait en ce sens moins risquée qu'une problématisation placée loin de l'origine : la première serait en effet – en vertu du narratif de GTT – basée sur des critères univoques et impliquerait un maigre enjeu social, tandis que la deuxième serait basée sur des critères équivoques et revêtirait un fort enjeu social.

9 Le cas de ces *ground truths* pour aider à la construction d'algorithmes dits de *voice analytics* a notamment été étudié dans Kang Edward, « Biometric imaginaries: Formatting voice, body, identity to data », *Social Studies of Science*, t. LIV, n°4, 2022.

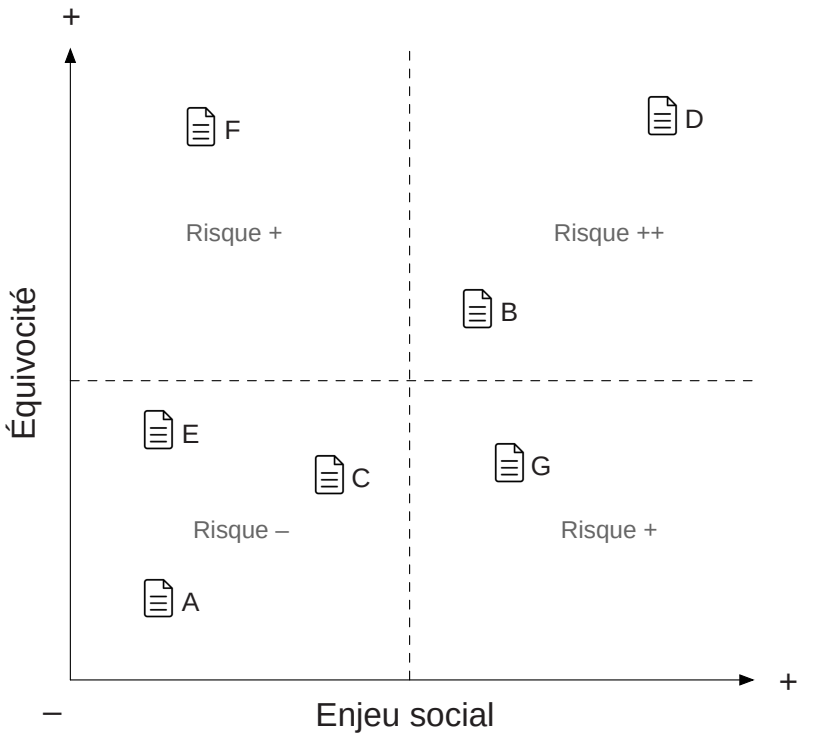


Figure 3 (ci-dessus). Tableau classificateur des problématisations algorithmiques en fonction de leur degré d'équivocité (ordonnée) et d'enjeux social (abscisse). Chaque point (icone document accompagné d'une lettre A, B, C, etc.) correspond à un compte rendu décrivant le processus de problématisation d'un projet d'algorithme donné. Une problématisation – telle que décrite dans un compte rendu dit de *ground-truth tracing* – est considérée d'autant plus risquée qu'elle cumule une forte équivocité et un fort enjeu social. Source : Librement inspiré de Kang Edward, « Ground truth tracings (GTT): On the epistemic limits of machine learning », op. cit., p. 8.

Imaginer que chaque *ground truth* mobilisée pour le développement algorithmique soit accompagnée d'un narratif de *ground-truth tracing* permettant de la positionner au sein d'un tableau qualificateur de risques (et donc de responsabilités et, éventuellement, de règles à respecter) peut paraître utopique, au mieux. Et pourtant, il s'agirait là sans doute d'un moyen de respecter l'importance des activités de problématisation algorithmiques en mettant celles et ceux qui s'y adonnent face à leurs responsabilités.

DIFFUSION–PERFORMANCE

Si documenter les processus de problématisation serait crucial pour mieux saisir les façons – plus ou moins risquées – dont sont cadrés les problèmes que les algorithmes sont amenés à résoudre, cela ne permettrait pas d'assurer leur innocuité technique et sociale. En effet, une fois les algorithmes effectivement construits – sur la base des *ground truths* qui leur servent de matrices fondamentales –, ceux-ci peuvent toujours produire des résultats biaisés et potentiellement néfastes[10]. D'où l'importance d'imaginer des procédés afin de s'assurer que les algorithmes – *après* avoir été usinés mais *avant* leur diffusion au sein de la société – produisent des résultats fiables et exempts de partis-pris injustifiés.

Plusieurs groupes de chercheur·euse·s ont proposé des moyens d'évaluer minutieusement les performances des algorithmes[11]. Mais l'un des plus intéressants – car ayant fait l'objet de tests empiriques – est sans doute celui proposé par Margaret Mitchell et ses collègues en 2019[12]. En s'inspirant notamment des organisations professionnelles de micro-électronique qui incitent les fabricants à joindre des fiches techniques certifiant des performances des composants cherchant à être commercialisés, Mitchell et ses collègues proposent une procédure de documentation qu'iels nomment *model cards*.

Ces *model cards* sont des documents de synthèse organisés en neuf sections, chacune se focalisant sur un aspect de l'algorithme à l'étude. Ces sections sont organisées comme suit :

10 C'est là un fait aujourd'hui bien établi : pour de nombreuses études rétrospectives sur les biais socio-culturels inscrits dans des algorithmes commerciaux utilisés pour la détection faciale, l'ajustement automatique de photographies, la justice pénale ou encore la détection de commentaires toxiques, les algorithmes sont des dispositifs propices à la perpétuation de préjugés fondamentalement injustes. Voir par exemple Buolamwini Joy. « How I'm fighting Bias in Algorithms », *TEDx*, 2016, URL : https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-63664, consulté le 9 juin 2019 ; Buolamwini Joy et Gebru Timnit, « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification », in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, PMLR, 2018 ; Dieterich William, Mendoza Christina, et Brennan Tim, « COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity », *Pro-Republica*, 8 juillet 2016, URL : https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html, consulté le 9 mai 2019 ; Dixon Lucas, Li John, Sorensen Jeffrey, Thain Nithum, and Vasserman Lucy, « Measuring and Mitigating Unintended Bias in Text Classification », in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New York, ACM, 2018.

11 Par exemple : Bender Emily M. et Friedman Batya, « Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science », *Transactions of the Association for Computational Linguistics*, t. VI, 2018 ; Holland Sarah, Hosny Ahmed, Newman Sarah, Joseph Joshua et Chmielinki Kasia, « The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards », 2018. Version préimprimée disponible à http://arxiv.org/abs/1805.03677, consulté le 9 juin 2020.

12 Mitchell Margaret, Wu Simone, Zaldivar Andrew, Barnes Parker et Vasserman Lucy, « Model Cards for Model Reporting », in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, ACM, 2019.

Détails du modèle (*model details*) : section qui vise à répondre aux questions de base vis-à-vis de l'algorithme, notamment qui l'a conçu ? Quand ? Avec quels financements ? Sous quel type de licence ? En se basant sur quel type d'architecture mathématico-informatique ?

Utilisation prévue (*intended use*) : section qui se penche sur les raisons d'être de l'algorithme et indique les domaines (e.g., divertissement, hôpital, science, enseignement) dans lesquels il convient de l'utiliser ou non.

Facteurs (*factors*) : section qui vise à lister les éléments capables d'impacter les performances de l'algorithme. Ces facteurs peuvent être *techniques*, en lien notamment avec les différents instruments (e.g., appareils photo, caméras) utilisés pour capturer les données que va traiter l'algorithme, ou encore les environnements (e.g., luminosité, humidité, pression) dans lesquels il va être déployé. Mais ces facteurs peuvent aussi être *sociaux*, en lien notamment avec les différentes populations visées par l'algorithme, certains groupes sociaux – notamment les plus marginalisés – pouvant être plus vulnérables que d'autres à un traitement algorithmique préjudiciable.

Métriques (*metrics*) : section qui rend compte et justifie des mesures utilisées afin d'évaluer les performances de l'algorithme : pourquoi telle mesure plutôt que telle autre ? Implique-t-elle des seuils de décision ? Et si oui, lesquels sont-ils et pourquoi ont-ils été choisis ?

Données d'évaluation (*evaluation data*) : section qui décrit les jeux de données mobilisés pour l'évaluation des performances : d'où proviennent ces jeux de données ? Qui les a assemblés ? Selon quelles procédures ? Quelles sont leurs limites ? Et pourquoi ont-ils été choisis ?

Données d'entraînement (*training data*) : section qui décrit les jeux de données mobilisées pour la construction de l'algorithme, c'est-à-dire la *ground truth* de laquelle il émane. Là encore, cela implique de répondre à des questions telles que : qui a construit la *ground truth* initiale ? Quand ? Selon quelle procédure ? Et est-elle libre d'accès ou propriétaire ?

Analyses quantitatives (*quantitative analysis*) : section qui présente les résultats de l'évaluation de l'algorithme en fonction des métriques indiquées précédemment, en indiquant si possible des valeurs de confiance.

Considérations éthiques (*ethical considerations*) : section qui rend compte des problématiques plus générales auxquelles est attaché l'algorithme. Cela peut concerner la sensibilité des données qu'il utilise (e.g., données confidentielles) ou encore les risques liés à son contexte d'utilisation (santé, sécurité).

Avertissements et recommandations (*caveats and recommandations*) : section conclusive qui liste les problèmes supplémentaires qui n'ont pas pu être inclus dans les sections précédentes. Par exemple, les résultats quantitatifs suggèrent-ils des tests supplémentaires ? Y a-t-il des groupes pertinents qui n'ont pas pu être représentés dans les jeux de données d'évaluation ? Si oui, comment serait-il possible de pallier ce manquement à l'avenir ?

À l'instar du *ground-truth tracing* proposé par Kang, les *model cards* proposées par Mitchell et ses collègues ont une ambition avant tout *descriptive* : elles visent à rendre explicites les apports et les limites des algorithmes. Mais à la différence du *ground-truth tracing* qui se positionne en amont de la construction effective des algorithmes (au moment de leur problématisation), les *model cards* se positionnent en aval de leur construction, juste avant leur diffusion potentielle au sein de la société en tant que produits informatiques.

Mais là encore, imaginer que chaque personne ou organisation souhaitant diffuser un nouvel algorithme soit pressée de soumettre un *model card* récapitulant ses performances peut paraître utopique, au mieux. Et pourtant, comme l'ont montré Mitchell et ses collègues dans leurs exemples empiriques (voir figure 4), cette documentation peut être à la fois suffisamment étoffée *et* suffisamment succincte pour permettre une évaluation approfondie sans engager d'encombrantes procédures administratives.

Figure 4 (page suivante). Exemple de *model card* pour le cas d'un algorithme de détection de sourire au sein d'images numériques. Si le travail sous-jacent à la présentation des performances et limites de l'algorithme est conséquent, son rendu est volontairement succinct, permettant dès lors une évaluation rapide (un peu à l'image des fiches techniques des composants électroniques). Source : reproduit avec permission à partir de Mitchell Margaret, Wu Simone, Zaldivar Andrew, Barnes Parker et Vasserman Lucy, « Model Cards for Model Reporting », op. cit, p. 7.

CONCLUSION

Qu'on le veuille ou non, les algorithmes impactent fortement nos vies quotidiennes. Au-delà de l'informatique mobile et des applications que nous sommes susceptibles de mobiliser (ou non), c'est toute une partie de l'infrastructure sociale (administration, droit, services de santé, banque) qui est de plus en plus irriguée par des méthodes informatiques de calcul. Mais ces algorithmes ne tombent pas du ciel : ils sont imaginés, définis et usinés dans de lieux spécifiques, par des personnes traversées de désirs qui mobilisent des outils dans lesquels sont inscrits des visions et des intérêts.

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

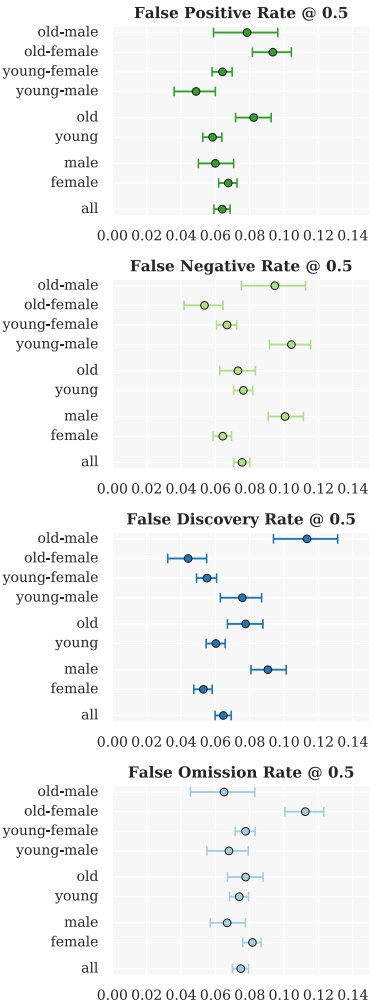
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

En ce sens, les algorithmes ne diffèrent en rien des autres dispositifs techniques : ils portent en eux – et dès lors reproduisent et promeuvent – les visions des agencements sociaux desquels ils émanent[13]. D'où l'importance de rendre explicites ces visions pour mieux fonder l'encadrement des algorithmes, quitte à réfréner quelque peu les ardeurs – le plus souvent commerciales – des évangélistes technologiques. C'est là, me semble-t-il, l'intérêt principal des réflexions que j'ai souhaité résumer ici : en sommant les constructeur·rice·s d'algorithmes à faire preuve de réflexivité (pour ne pas dire d'humilité), les procédures de *ground-truth tracing* ou de *model cards* pourraient insuffler un principe de précaution à même de rendre le développement algorithmique digne des bouleversements qu'il provoque.

13 Voir à ce sujet le texte fondamental de Akrich Madeleine, « La description des objets techniques », in Callon Michel et Latour Bruno (dir.) *Sociologie de la traduction : Textes fondateurs*, Paris, Presses des Mines, 2013.