# Les algorithmes d'aide à la décision dans l'aide sociale à l'enfance : interroger la cohérence des outils d'une collectivité locale à l'autre au Royaume-Uni

**Thomas M. Vogl**
Doctorante en information, communication, et les sciences sociales, Oxford Internet Institute, Université d'Oxford, Royaume-Uni

## Introduction

La présente contribution soulève un problème relatif à la compatibilité des outils algorithmiques d'aide à la décision dans l'aide sociale à l'enfance au sein des collectivités locales du Royaume-Uni, bien que l'énoncé du problème et les résultats aient aussi un impact qui s'étend à d'autres administrations et secteurs qui utilisent ces mêmes outils. Elle est également pertinente dans les discussions à propos de l'incohérence, de la standardisation et de la coordination entre les entités administratives. L'argument principal ici développé est que l'implémentation de nouveaux outils d'algorithmes d'aide à la décision dans le secteur public pourrait introduire un genre différent d'incohérences dans le processus de prise de décision et que nous pourrions manquer d'informations pour déterminer si c'est bien le cas.

Lors de la dernière décennie, les collectivités locales au Royaume-Uni ont subi d'importantes mesures d'austérité (Atkins 2020; Financial Times 2019; Harris, Hodge, and Phillips 2019). En réponse, elles ont mis en place des stratégies pour augmenter le rendement et l'efficacité de leurs services, y compris au moyen d'outils d'analyse prédictive pour aider à la prise de décision (Bright et al. 2019; Dencik et al. 2018; Symons 2016b). Les collectivités locales ont déployé ces outils sur l'ensemble de leur gamme de services, y compris sur des sujets sensibles comme l'aide **(p.271)** sociale à l'enfance (Redden, Dencik, and Warne 2020; Vogl 2021), et elles travaillent auprès de nombreux organismes pour développer ces outils (Booth 2019; Sena et al. 2019; Turner 2019b). Alors qu'il existe des discussions éthiques sur l'utilisation des outils d'analyse prédictive pour aider à la prise de décision dans le champ de ce service (Clayton 2019; ICO 2019; Leslie et al. 2020), on en sait peu sur les répercussions de l'incohérence entre les différents outils utilisés par les collectivités locales.

Cette étude analyse des documents afin d'identifier les usages des analyses prévisionnelles utilisant l'apprentissage automatique dans l'aide sociale à l'enfance, comprendre les caractéristiques de ces modèles, et soulever le problème de la compatibilité des outils lorsque plusieurs fournisseurs sont impliqués dans le développement des outils d'aide algorithmique à la décision et qu'une multiplicité de ces outils existe au sein d'un même secteur. Elle prend ensuite les prêts hypothécaires en exemple pour démontrer que précision du modèle et équité ne sont pas nécessairement liées, puis elle explore une hypothèse pour mettre en lumière les différences de décisions recommandées, pour un même individu au sein du secteur de l'aide sociale à l'enfance, en fonction du modèle sélectionné par une collectivité territoriale donnée.

Les résultats suggèrent une incohérence entre ces outils, et que (toutes choses égales par ailleurs) ces différences peuvent amener à des recommandations variées pour le même individu selon le modèle utilisé par la collectivité territoriale ; mais, plus important, qu'il y a un manque d'information publique concernant les modèles utilisés dans le secteur de la protection de l'enfance[1] pour évaluer efficacement la cohérence des outils actuellement en service. Ces résultats suggèrent que plus d'attention doit être accordée non seulement à la cohérence des outils entre les organismes proposant des services similaires sur différents territoires, mais également à une transparence suffisante pour pouvoir juger de la compatibilité ou de l'incompatibilité des outils utilisés par différents organismes. **(p.272)**

Dans ce qui suit, la partie « Contexte » met en place la discussion sur le sujet, la toile de fond et la problématique. L'analyse de la littérature se penche sur la confidentialité de la prise de décision dans le secteur public, l'introduction d'outils algorithmiques, et les biais et l'équité dans le domaine de l'apprentissage automatique. La partie « Méthodologie » décrit comment j'ai rassemblé, ou échoué à rassembler, les informations qui alimentent la conclusion. La partie « Résultats » présente les découvertes à propos des différences de modèles entre les collectivités locales du Royaume-Uni, les problèmes de transparence, et les risques de prise d'une décision non pertinente, illustrés par le cas pratique de la modélisation prédictive des risques dans les prêts hypothécaires. La partie « Discussion » met en avant les implications des résultats avant que l'article conclue avec les éléments clés à retenir et propose des ouvertures sur de nouveaux champs de recherche.

A – Contexte

En arrière-plan de cette recherche, se trouve l'intérêt grandissant pour les outils d'analyse prévisionnelle par les autorités publiques, qui les adoptent pour assister leurs prises de décision. Les autorités locales du Royaume-Uni (et à l'international) s'approprient différents algorithmes d'analyse prévisionnelle (Bright et al. 2019; Misuraca and van Noordt 2020). Un nombre en constante augmentation d'autorités locales se servent, d'une façon ou d'une autre, d'analyses prévisionnelles dans leurs services sociaux (Marsh and McIntyre 2020). Beaucoup de ces algorithmes concernent l'aide sociale à l'enfance (Turner 2019a). La principale restriction de ces développements technologiques au Royaume-Uni est qu'il n'y a pas de processus standard pour assurer la transparence de ces outils prédictifs (Dencik et al. 2018), bien que certains organismes aient commencé à militer pour une plus grande transparence (Ada Lovelace Institute 2020; Centre for Data Ethics and Innovation 2019; Vogl 2021), et que d'autres pays, comme la France, aient rendu obligatoire la transparence des algorithmes publics et commencé à orienter le développement des registres (Pénicaud 2021). **(p.273)**

La croissance de l'analyse prévisionnelle intervient dans le contexte d'une longue tradition publique de collecte de données en tant que composante de la prestation du

[1] Bien que certains organismes se soient engagés à une transparence totale concernant leur modèle et leur approche (Clayton, Gibbons, and Sanders 2019).

service. L'une des préoccupations des États est l'identification précise des résidents dans leurs périmètres géographiques (Alonso and Starr 1987) et l'utilisation de ces données nominatives, soit pour des raisons de simplification et de lisibilité (Scott 1998), soit de soutien de la gestion administrative de la société (Weber 1968), soit de fluidification des mesures (Foucault 1991). Ces activités de simplification et de mesure ont contribué au respect d'une égalité par la procédure des prestations de services (Weber 1968), malgré quelques différences sur le plan de la réalité de terrain (Lipsky 2010; Zacka 2017).

Aujourd'hui, de nombreux outils peuvent aider à dégager des aperçus de données, structurées ou non, et d'aucuns en proposent l'emploi dans la prise de décision (Mayer-Schönberger and Cukier 2013; Pentland 2014). D'autres secteurs de l'administration publique tentent d'utiliser ces outils à leur avantage, et des aperçus qu'ils peuvent fournir, pour améliorer la prise de décision (Dunleavy et al. 2006; Kim, Trimi, and Chung 2014; Toderas and Manning 2017). Cependant, ces efforts ont révélé que, bien que les mesures constituent une partie importante des efforts de l'État, les formes que prennent les données et les registres ainsi que les compétences privilégiées par le secteur public ont élevé des barrières à l'assimilation des technologies numériques (Dunleavy et al. 2006; Pollitt 2009). De nombreux problèmes ont été identifiés en tant qu'obstacles à la mise en place heureuse de ce genre de projets de gouvernance digitale, tels que ceux en lien avec le traitement et l'analyse des données dans le cadre de l'aide automatisée à la décision (Yang and Maxwell 2011). Une barrière, relative à l'utilisation des technologies pour l'analyse des données administratives, a été attribuée à des limites de capacité interne à la fois dans la mise à disposition des données et dans la réalisation des analyses (Allard et al. 2018). En conséquence, les organismes ont parfois accès à des éléments **(p.274)** d'information différents et disposent de capacités internes inégales, ce qui peut impliquer des partenariats avec d'autres organismes (Bright et al. 2019).

Dans le contexte de la transformation numérique au sein des autorités locales du Royaume-Uni, dont de nombreux services sont opérés au niveau local et où beaucoup du travail quotidien lié à leur prestation et leur administration sont décentralisés, les autorités locales sont libres de choisir leur propre méthode pour identifier les outils adéquats à l'assistance à la prise de décision (Malomo and Sena 2017). Beaucoup guettent les occasions de tirer parti de leurs données et des analyses avancées de leur portfolio de services (Bright et al. 2019; Dencik et al. 2018; M. Eaton and Bertoncin 2018; Symons 2016a; 2016b). La variabilité des données et des capacités mène à des différences dans les technologies qu'elles adoptent. Certains s'enorgueillissent de leur capacité interne à grand renfort de déclarations telles que : « Nous avons la chance que les technologies de l'information ne soient pas une barrière pour nous, car nous avons notre propre équipe technique... Nous disons "il nous faut ceci", ils le comprennent et nous travaillons ensemble pour atteindre cet objectif. » (Bright et al. 2019, 31) Mais d'autres doivent trouver ce savoir-faire auprès de sources externes, une démarche que certains observateurs commentent ainsi : « Au niveau local de Gouvernement, je dirais qu'ils n'ont habituellement pas la capacité de développer par eux-mêmes l'apprentissage automatique, mais qu'ils ont une plus grande marge de manœuvre sur

la façon dont ils l'implémentent. » (Bright et al. 2019, 30) Ces différences peuvent mener à des incompatibilités entre les manières dont les organismes recherchent, développent, implémentent et évaluent les projets d'algorithmes d'aide à la décision.

L'aide sociale à l'enfance est un des domaines dans lesquels les technologies d'analyse sont recherchées (Garrett 2005; Gillingham 2019; Hall et al. 2010). Ce secteur requiert un rythme soutenu dans la prise de décision, dans un environnement aux enjeux importants (Sanders et al. 2017). Par exemple, lorsque les décisions ont pour conséquence le placement ou non d'un enfant, et lorsque la conséquence de ces décisions peut mener **(p.275)** soit à la séparation de la famille ou, dans le pire des cas, à la mort d'un enfant (Lord Laming 2003; LSCB Haringey 2009). Au Royaume-Uni, ce service est réglementé au niveau national, mais sa prestation s'effectue localement par le biais d'environ deux cents autorités réparties dans le pays. Dans l'aide sociale à l'enfance, en particulier, les autorités locales testent et adoptent un certain nombre d'options d'apprentissage automatique afin de répondre à leurs besoins de modélisation prédictive des risques (Booth 2019; Celia 2016; Sanders et al. 2017; Sena et al. 2019; WW for CSC 2019). Cependant, la pluralité des approches entre les collectivités locales soulève d'importantes questions d'équité, notamment en ce qui concerne la cohérence des décisions. Si la diversité des options aide les autorités locales à trouver individuellement des solutions qui répondent à leurs besoins, elle a pour conséquence que des outils différents sont développés et utilisés par ces autorités. Le risque est de conduire à une prise de décision sensiblement différente d'une autorité locale à l'autre.

L'étude de la modélisation prédictive des risques au niveau local présente trois avantages principaux : elle souligne la nature locale et contextuelle des projets technologiques ; elle met en évidence le nombre de collectivités locales susceptibles de prendre des décisions indépendantes concernant l'adoption d'algorithmes ; et permet ainsi d'illustrer les différences potentielles entre ces algorithmes.

B – Analyse de la littérature

Le problème abordé dans cette étude est lié à quatre débats clés : l'équité sociale et l'égalité procédurale dans l'administration publique ; la juste place de l'apprentissage automatique dans les processus de prise de décision ; l'équité, la responsabilité et la transparence dans l'apprentissage automatique ; et la pertinence du modèle dans un but d'exactitude et d'équité.

Les principaux points importants pour l'argumentaire de ce document dans chacune de ces questions sont les suivants :

- dans le secteur public, l'exercice de la discrétion professionnelle peut introduire des incohérences (Durose 2011; Lipsky 2010) et ce phénomène a été identifié dans la protection de l'enfance (Fluke et al. 2014; Keddell 2014) ; **(p.276)**
- on estime que les algorithmes réduisent les incohérences dues à la discrétion des agents, même lorsque les décisions sont toujours prises par des humains (Chouldechova et al. 2018). Cependant, lorsque les décisions finales sont toujours prises par des humains, il a été constaté que les outils d'aide à la

décision exercent une influence sur la responsabilité et la prise de décision (Green and Chen 2019; Snow 2021) ;
- en raison de l'influence que les algorithmes peuvent avoir sur la prise de décision, il existe des préoccupations concernant la partialité et l'équité, découlant de la nature de « boîte noire » de certains algorithmes et du défi que cela représente pour la transparence publique (Pasquale 2015) ;
- cependant, tout en diminuant potentiellement la discrétion, les algorithmes peuvent introduire de nouveaux types d'incohérences en fonction de leurs caractéristiques (par exemple, les entrées de données, le modèle ou ses spécifications) (Lee and Floridi 2021).

La question de la cohérence des outils d'aide à la décision par apprentissage automatique dans les collectivités locales peut être rattachée à des idées plus globales d'équité dans la théorie des contrats sociaux et l'étude de l'administration publique. Rawls, dans sa théorie de la justice en tant qu'équité, a proposé que si nous nous trouvions tous dans une position initiale hypothétique derrière un « voile d'ignorance » sur nos caractéristiques personnelles, y compris celles liées à la géographie, à la démographie et au statut socio-économique, nous choisirions des règles pour la société qui mènent à des résultats relativement équitables pour tous (Rawls 2009). De même, dans l'évolution des systèmes administratifs, Weber a observé qu'en passant du patrimonialisme à l'organisation bureaucratique, la manière dont les décisions étaient prises passait du favoritisme à l'égalité procédurale fondée sur des règles (Weber 1968). Plus récemment, des chercheurs se sont penchés sur la numérisation croissante des services publics et ont suggéré que le traitement et l'analyse des données à l'aide des technologies de l'information peuvent faire passer l'administration publique d'une approche fondée sur l'égalité procédurale à une approche fondée sur l'égalité des résultats (Dunleavy and Margetts 2015). Dans cette logique, d'autres chercheurs s'intéressent aux multiples façons dont la numérisation et les algorithmes modifient la nature de l'administration **(p.277)** publique (Bovens and Zouridis 2002; Lorenz, Meijer, and Schuppan 2021; Vogl et al. 2020). Qu'il s'agisse des principes, des règles ou des résultats, on est en droit d'en attendre une certaine cohérence au-delà des frontières régionales. Par exemple, personne ne souhaiterait que des décisions importantes concernant l'enquête ou le retrait des enfants de leur famille et leur prise en charge dépendent du modèle particulier choisi par une autorité locale donnée. Dans la pratique, il se peut que nous ne nous trouvions pas derrière un voile d'ignorance, mais plutôt face à une boîte noire qui fait disparaître la possibilité de transparence pour ces différents outils (Kolkman 2020; Vogl 2021). Le point essentiel est que l'introduction d'outils technologiques ne conduit pas de façon systémique à des règles procédurales équitables ou à l'égalité des résultats.

Les travaux académiques sur la prise de décision en matière de protection de l'enfance et sur l'influence exercée par les technologies de l'information ne datent pas d'hier et restent pertinents en raison de l'intérêt contemporain pour la juste place de l'apprentissage automatique lors des processus décisionnels dans la protection de l'enfance. Des recherches antérieures ont examiné les incohérences intrinsèques à la mise en œuvre par les autorités locales de la politique nationale en matière de

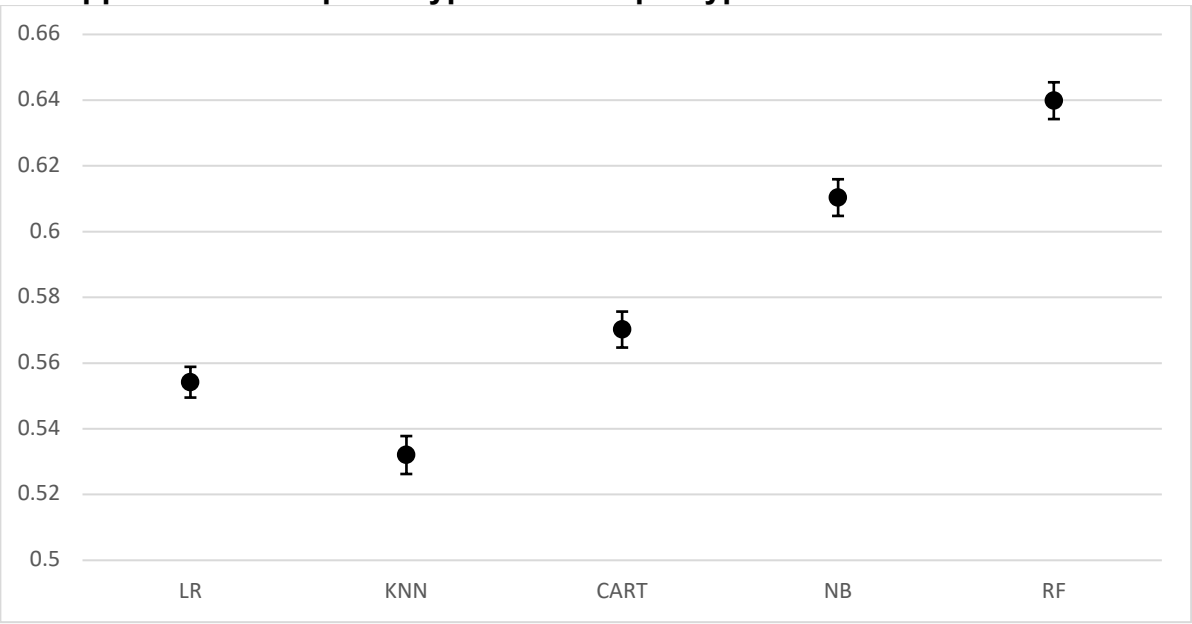
technologies de l'information dans l'aide sociale à l'enfance (Peckover, Hall, and White 2009). Des recherches plus récentes sur le potentiel de l'analyse prédictive se montrent plus optimistes. Certains chercheurs examinent les possibilités offertes par la modélisation prédictive des risques dans les services de protection de l'enfance (Cuccaro-Alamin et al. 2017) et cherchent à améliorer l'efficacité générale de ces modèles (Coston et al. 2020). D'autres remettent en question l'efficacité de ces outils (Clayton et al. 2020), bien qu'ils reconnaissent qu'elle a pu être affectée par les limitations des données, ce que d'autres chercheurs ont désigné comme un obstacle au développement de modèles efficaces (Cuccaro-Alamin et al. 2017; Shafiq 2020). Il existe également des recherches sur l'équilibre adéquat entre la prise de **(p.278)** décision par la machine et la prise de décision par l'humain, qui examinent dans quelle mesure les machines peuvent orienter efficacement les décisions d'ordre de priorité et à quel moment elles deviennent moins précises, nécessitant l'intervention de décideurs humains (Raghu et al. 2019). Certes, ces derniers ne sont pas parfaits et peuvent également être partiaux et incohérents (Fluke et al. 2014; Keddell 2014), mais nous sommes en général plus à l'aise avec l'intervention humaine dans ces cas limites (Bigman and Gray 2018), quand bien même l'intervention humaine pourrait réduire l'efficacité des prédictions (Green and Chen 2019), échouer à résoudre les problèmes sous-jacents du modèle (Goldenfein 2019) ou être tout aussi opaque (Zerilli et al. 2019).

Les préoccupations relatives à la partialité et à l'équité commencent à s'inviter de manière plus formelle dans le domaine de la recherche sur l'apprentissage automatique, comme en témoigne l'apparition récente d'une conférence spécifique sur l'équité, la responsabilité et la transparence dans l'apprentissage automatique (ACM FAccT Conference 2018). La recherche sur ces thèmes se concentre généralement sur les préoccupations liées à la partialité des modèles d'apprentissage automatique dans les cas individuels et à l'équité des prédictions affectant des groupes au cas par cas (Eubanks 2017; Kearns et al. 2019; Obermeyer and Mullainathan 2019; Verma and Rubin 2018), la capacité des algorithmes d'apprentissage automatique à révéler les biais (Benthall and Haynes 2019; Madras et al. 2019), la transparence des opérations de ces modèles (Munro 2019; Pasquale 2015), la possibilité de l'explicabilité par un apprentissage automatique interprétable (Wachter, Mittelstadt, and Russell 2018; Watson 2020), ainsi que la conception et le développement d'algorithmes (Janssen and Kuk 2016). Bien qu'il y ait également eu une **(p.279)** exploration récente de la façon dont divers algorithmes peuvent mieux répondre à différentes définitions de l'équité, ces débats ne comportent pas de discussion sur la possibilité d'expliquer l'équité (Lee 2019). Dans ces débats, il manque une discussion sur l'équité d'une décision pour un individu précis et, en particulier, sur la cohérence de cette décision entre plusieurs modèles. Cette lacune est particulièrement préoccupante dans le cadre des débats sur l'utilisation de l'apprentissage automatique pour l'analyse prédictive dans l'aide sociale à l'enfance.

Un article particulièrement pertinent dans le domaine de l'éthique et de l'apprentissage automatique porte sur les compromis en matière d'équité dans l'utilisation de différents types de modèles d'apprentissage automatique (Lee and Floridi 2021). Cet article est important non seulement parce qu'il met en évidence les compromis éthiques que les

décideurs doivent prendre en compte lors de la sélection d'un modèle d'apprentissage automatique, mais aussi, dans le cadre de la présente recherche, en raison de ce que ces compromis peuvent nous révéler sur les décisions sur le plan individuel lorsque plusieurs types de modèles différents sont utilisés. L'article explore les compromis d'équité à travers un cas d'utilisation dans le domaine des prêts hypothécaires aux États-Unis, en utilisant des données recueillies en application du Home Mortgage Disclosure Act et en mesurant l'équité au sein de deux groupes : les Noirs et les Blancs non hispaniques (Lee and Floridi 2021). Les résultats suggèrent que la précision du modèle et l'équité ne sont pas nécessairement liées.

**Précision des modèles d'apprentissage automatique dans la prédiction des approbations de prêts hypothécaires par type de modèle**




Source : Figure adaptée de Lee et Floridi, (2021, 173). **(p.280)**

La figure ci-dessus expose comment différents modèles se comportent sur un ensemble de données de prêts hypothécaires approuvés ou refusés pour deviner quelle réponse a été donnée pour chacun. Les modèles comprennent la régression logistique (LR), les K plus proches voisins (KNN), un arbre de classification et de régression (CART), une classification naïve bayésienne (NB) et une forêt aléatoire (RF). Le point à retenir est qu'ils diffèrent en termes de précision. Ces modèles sont ensuite comparés selon trois mesures d'équité, et le résultat démontre que les modèles les plus performants diffèrent selon la mesure choisie et que les plus précis ne sont pas nécessairement les plus équitables. Les mesures d'équité dans cet exemple particulier sont l'égalité des chances (qui évalue les algorithmes en fonction de l'équilibre entre les groupes sur les taux de faux négatifs, faussement approuvés, et de faux positifs, faussement refusés), la parité prédictive positive (qui évalue les algorithmes en fonction de la probabilité qu'ils prédisent le remboursement parmi les demandeurs solvables), et l'équilibre positif des classes (qui évalue les algorithmes en fonction du score de probabilité moyen des demandeurs noirs et blancs solvables qui remboursent leurs prêts) (Lee and Floridi 2021). Le tableau ci-dessous illustre comment le niveau d'équité d'un modèle dépend de la mesure d'équité choisie.

**Pourcentage de disparité entre deux groupes sur différentes mesures d'équité par type de modèle**

| Modèle | Egalité des chances | Parité predictive positive | Équilibre positif de la classe |
|---|---|---|---|
| **LR** | 4% | 33% | 2% |
| **KNN** | 6% | 18% | 8% |
| **CART** | 6% | 21% | 16% |
| **NB** | 7% | 44% | 41% |
| **RF** | 9% | 3% | 18% |

Source : Tableau adapté de Lee et Floridi, (2021, 176).

Cela suggère que certains individus pourraient être traités différemment en fonction du modèle choisi et pas seulement en raison de la précision. Les pourcentages représentent le degré de disparité et indiquent à la fois la disparité la plus faible (surlignée en bleu) et la disparité la plus forte (surlignée en orange). Une disparité plus élevée représente une plus grande violation des conditions de la mesure d'équité.

Nous pouvons voir dans le tableau 1 que même s'il existe une hiérarchie dans la précision, chacun de ces modèles renvoie des résultats différents selon les mesures d'équité. Cela suggère non seulement que lors de la sélection d'un **(p.281)** modèle, nous devrions nous concentrer sur autre chose que la précision, mais aussi que ces modèles procurent différentes prédictions, qu'elles soient correctes ou non, sur le plan individuel, indépendamment de leur précision relative. L'élément intéressant ici est la possibilité que les personnes dans les catégories de prédictions correctes et incorrectes puissent changer d'un modèle à l'autre. Par exemple, une demande de prêt correctement prédite comme acceptée par un modèle de régression logistique pourrait être refusée par erreur par un modèle de forêt aléatoire, bien que le modèle de forêt aléatoire soit globalement plus précis. Alors qu'auparavant, les algorithmes gouvernementaux pouvaient être des processus fondés sur des règles, dans lesquels tout individu répondant aux critères d'éligibilité était traité d'une façon particulière, le risque est que de tels algorithmes puissent renvoyer des résultats différents pour le même individu en fonction du type de modèle utilisé.

La présente recherche se concentre spécifiquement sur l'impact potentiel, sur les décisions au niveau individuel, de l'utilisation de différents modèles par différents prestataires et sur l'existence éventuelle d'une incohérence entre ces modèles. Si des différences de résultats sont constatées, cela soulève des questions quant à l'influence de l'adoption de plusieurs modèles sur l'équité et la cohérence de la prise de décision entre les différents prestataires des autorités locales, et cela fournit une justification supplémentaire pour une plus grande transparence algorithmique et un examen continu de l'organisation sociotechnique de la prise de décision algorithmique dans l'aide sociale à l'enfance, dans un effort pour soutenir une plus grande cohérence de la décision entre les autorités locales.

C – Méthodologie

Cette étude examine les différences, dans les caractéristiques et les approches, des utilisations de l'analyse prédictive dans les services sociaux de l'enfance, puis démontre les impacts potentiels de ces différences sur la cohérence des décisions sur le plan individuel. Elle compare d'abord des cas clés dans le contexte britannique afin d'explorer et de mieux comprendre les similitudes et les différences entre le développement d'outils par chacun des types d'organismes qui sont impliqués dans leur conception. Dans ce contexte, il existe cinq grandes catégories de fournisseurs de ces outils : les organismes publics non ministériels (WW for CSC 2018), les sociétés anonymes à but social (Sanders et al. 2017), les universités (University of Essex 2019), les collectivités locales elles-mêmes (Booth 2019) **(p.282)** et les entreprises privées (Marsh 2019; Xantura 2016). Ce document se concentre sur les projets de cinq organismes pour illustrer ces catégories : le travail de What Works for Children's Social Care, la Behavioural Insights Team, le Catalyst Project de l'université d'Essex, le centre d'analyse interne de Bristol et le Early Help Profiling System de Xantura. Bien que d'autres organisations que celles listées soient également impliquées dans l'analyse prédictive au niveau local, ces cinq cas ont été sélectionnés car ils représentent les principaux groupes organisationnels impliqués dans le développement d'outils prédictifs.

Cette comparaison permet de se concentrer sur les caractéristiques particulières des cas qui présentent un intérêt et clarifie également les principales différences de variables et de spécifications entre les modèles, ainsi que les cas où les informations pertinentes ne sont pas disponibles (Tsoukas 2009). Bien que la comparaison se concentre sur les détails contextuels associés à chaque cas, elle permet également de s'éloigner des détails spécifiques de ces cas particuliers pour s'intéresser à la structure plus profonde du phénomène, dans le but d'expliquer comment le développement d'outils dans différentes organisations au sein d'un même secteur de services peut avoir un impact sur la cohérence de la prise de décision à travers les frontières juridictionnelles régionales de ce secteur (Flyvbjerg 2006; Orlikowski and Baroudi 1991).

Pour étayer son argumentation, l'étude s'est appuyée sur l'analyse de documents et sur une expérience de pensée adaptée de l'exemple des prêts hypothécaires. L'analyse de documents a été utilisée pour examiner les informations accessibles au public sur les différentes organisations qui engagent ces projets et pour synthétiser les détails de leurs outils (Thorne et al. 2004; Timulak 2014). Il est important de noter qu'il n'existe pas de registre central de l'utilisation des outils algorithmiques dans le secteur public britannique. La recherche s'est donc appuyée sur l'analyse documentaire pour parcourir la littérature journalistique, grise et savante afin d'identifier les types d'organisations développant des algorithmes, les types et les détails de ces algorithmes et le niveau de transparence les entourant, dans les quelque deux cents autorités locales du Royaume-Uni fournissant des services d'aide sociale à l'enfance. **(p.283)**

Les documents analysés comprenaient des rapports de médias et d'organisations professionnelles (BASW 2019; Booth 2019; L. Eaton 2019; Graham 2017; Shafiq 2018;

Sheridan 2018; SocialWhatNow 2019; Stevenson 2018; Turner 2019a; 2019b), des rapports et des protocoles publiés par les organisations qui développent les outils (Celia 2016; Clayton 2019; Sanders et al. 2017; Toderas and Manning 2017) et des rapports d'organismes assurant une surveillance indépendante, comme le Commissariat à l'information (ICO 2019; Leslie et al. 2020). L'étude a également examiné les recherches secondaires sur l'utilisation de l'analyse prédictive dans les collectivités locales (Bright et al. 2019; Dencik et al. 2018; Symons 2016b; Vogl 2021). L'analyse documentaire a été utilisée pour définir le contexte, identifier les caractéristiques des outils et les différences entre eux, affiner les limites du site et se familiariser avec les préoccupations, les concepts et la terminologie utilisés par les différents groupes (Benbasat, Goldstein, and Mead 1987).

L'analyse documentaire a également été utilisée de manière à aller au-delà du contenu des documents pour comprendre le contexte, plus précisément s'il était possible d'identifier si ces types de projets sont suffisamment transparents pour permettre le niveau d'examen nécessaire afin de déterminer s'il existe des problèmes de cohérence des modèles (Karppinen and Moe 2012). Dans l'ensemble, l'objectif de l'analyse documentaire était de déterminer quels modèles étaient utilisés dans les collectivités locales, s'ils différaient, et dans quelle mesure les informations étaient suffisamment transparentes pour identifier où ils étaient utilisés, ainsi que leurs caractéristiques.

La seconde méthode utilisée est une expérience de pensée basée sur les incohérences potentielles, découlant de leurs performances divergentes en termes de précision et d'équité, entre les modèles de prêt hypothécaire. L'expérience de pensée transpose ces résultats au cas de la modélisation **(p.284)** prédictive des risques dans le domaine de l'aide sociale à l'enfance afin d'illustrer les possibles enjeux pour la cohérence de la prise de décision au niveau individuel entre les autorités locales utilisant différents modèles d'apprentissage automatique. Cette expérience de pensée est destinée à fournir une motivation théorique pour le problème de la cohérence des modèles dans un secteur qui en utilisent plusieurs.

D – Résultats

Les résultats indiquent qu'il existe différents modèles utilisés par les autorités locales, que les approches primaires dans leur conception diffèrent selon les organisations impliquées, que ces différents modèles pourraient générer des résultats incohérents pour les mêmes individus et, enfin, que dans certains cas, les informations disponibles au public sont insuffisantes. Cette section illustre ces points par une description des informations disponibles sur les modèles utilisés par les collectivités locales, l'expérience de pensée sur la cohérence des modèles et un commentaire général sur la transparence.

Dans les cas où des informations étaient disponibles, trois types de modèles différents se sont avérés être les plus utilisés dans le secteur : la forêt aléatoire, la régression logistique et le *boosting* par gradient. En général, lorsque les autorités locales travaillaient avec des organismes publics non ministériels, des sociétés à vocation

sociale ou des universités, il y avait un niveau de transparence plus élevé autour des modèles et des variables utilisés. La Behavioural Insights Team a utilisé une combinaison de données administratives structurées et de traitement du langage naturel des dossiers des travailleurs, en utilisant la modélisation thématique structurelle, dans le cadre d'un algorithme d'apprentissage automatique (Sanders et al. 2017), tandis que l'université d'Essex s'est appuyée uniquement sur un ensemble de variables dérivées des données administratives, en appliquant les méthodes du lasso et de la forêt aléatoire en langage R (Sena et al. 2019). Le What Works Centre for Children's Social Care a fait preuve de la plus grande transparence, en définissant ses variables indépendantes et dépendantes, ainsi que les spécifications du modèle, et en indiquant que tout le code serait publié sous une licence copyleft sur GitHub à la fin du projet (Clayton, Gibbons, and Sanders 2019). Lorsque les collectivités locales travaillaient avec des entreprises privées, moins d'informations sur les modèles et les variables étaient disponibles. Dans le tableau qui suit, les cinq organisations clés examinées dans cette étude sont présentées avec des informations sur leurs modèles, leurs variables et les prédictions recherchées. **(p.285)**

**Modèles et variables par organisation**

| Organisation | Modèles | Variables indépendantes | Variable dépendantes |
|---|---|---|---|
| **Organismes publics non ministériels** (par exemple, What Works for Children's Social Care (Clayton, Gibbons, and Sanders 2019)) | - Arbre de décision<br>- Régression logistique<br>- *Boosting* par gradient | - Données administratives quantitatives de l'aide sociale à l'enfance<br>- Modélisation thématique structurelle des dossiers | - Réorientations<br>- Escalades |
| **Sociétés anonymes à finalité sociale** (par exemple, Behavioural Insights Team (Sanders et al. 2017)) | - Arbre de décision<br>- *Boosting* par gradient | - Données administratives quantitatives de l'aide sociale à l'enfance<br>- Modélisation thématique structurelle des dossiers | - Réorientations<br>- Escalades |
| **Universités** (par exemple, Catalyst Project, University of Essex (Sena et al. 2019)) | - Régression linéaire<br>- Régression logistique<br>- Arbre de décision<br>- Forêt aléatoire | - Données administratives quantitatives de l'aide sociale à l'enfance | - Risque de prise en charge<br>- Examen par un MASH (Multi-Agency Safeguarding Hub) |
| **Les autorités locales** (par exemple, le conseil municipal de Bristol (Booth | - IBM SPSS pour la modélisation prédictive | - Données administratives provenant de l'ensemble des autorités locales (par exemple, éducation, | - Risque d'exploitation des enfants |

| | | | |
|---|---|---|---|
| 2019; Dencik et al. 2018)) | | justice pénale, allocations, santé) | |
| **Entreprises privées** (par exemple, Xantura (Graham 2017; McIntyre and Pegg 2018; Sheridan 2018; Stevenson 2018; Xantura 2018)) | - Indisponible | - Variables exactes indisponibles | - Risque d'abus |

Sources : Clayton et al., (2019); Sanders et al., (2017); Sena et al., (2019); Booth, (2019); Dencik et al., (2018); Graham, (2017); McIntyre et Pegg, (2018); Sheridan, (2018); Stevenson, (2018); Xantura, (2018). **(p.286)**

Ce tableau résume les résultats de l'analyse documentaire. Dans la colonne organisation figurent les cinq principales catégories de fournisseurs de ces outils, précédemment exposées dans la partie « Méthodologie » : les organismes publics non ministériels (WW for CSC 2018), les sociétés anonymes à vocation sociale (Sanders et al. 2017), les universités (University of Essex 2019), les collectivités locales elles-mêmes (Booth 2019) et les entreprises privées (Xantura 2016). La catégorie des entreprises privées montre moins de transparence sur les types de modèles et les données utilisées. Le même manque de transparence est présent en ce qui concerne les autorités locales lorsqu'elles effectuent leur propre modélisation des risques en s'appuyant sur des logiciels privés pour leurs outils de modélisation. Dans certains cas, ce manque de transparence est justifié par le fait que ces informations sont commercialement protégées (Dencik et al. 2018). Malgré ces lacunes dans les informations, lorsque les données sont disponibles, le tableau montre qu'il existe des différences dans les modèles, les variables et les entrées de données.

En revenant à l'exemple du prêt hypothécaire et aux compromis entre la précision et l'équité, nous pouvons voir non seulement comment les décisions sur la sélection du modèle deviennent plus compliquées, mais aussi comment la divergence entre précision et équité révèle un fait important sur la cohérence du modèle pour des individus particuliers. Elle peut aider à illustrer les conséquences des différences de modèles sur le plan individuel. Bien que le modèle de forêt aléatoire soit plus précis, il est moins équitable sur certaines mesures. Cela signifie que, malgré son efficacité dans l'ensemble, il est moins précis pour des sous-groupes particuliers que la régression logistique. Sur la base de ces deux facteurs (précision et équité), je souhaite proposer une expérience de pensée qui va au-delà de la question de savoir s'ils diffèrent uniquement sur ces points. Ce qui est plus intéressant, c'est la possibilité que les personnes appartenant aux catégories correctement et incorrectement prédites puissent changer d'un modèle à l'autre. Comme nous l'avons mentionné précédemment, une demande de prêt correctement prédite dans le cadre d'un modèle de régression logistique pourrait être refusée à tort dans le cadre d'un modèle **(p.287)**

de forêt aléatoire, bien que le modèle de forêt aléatoire ait un niveau de précision plus élevé.

L'extension de l'exemple des prêts hypothécaires ci-dessus au contexte de la protection de l'enfance permet de développer une proposition théorique sur la cohérence des décisions. De manière pratique, nous pouvons également constater que la régression logistique et la forêt aléatoire ont été utilisées à la fois dans l'exemple des prêts hypothécaires et dans celui de la modélisation du risque prédictif en protection de l'enfance, ce qui permet de relier les deux cas. Malgré la différence de contexte, on peut supposer que le modèle de forêt aléatoire reste le plus précis. Nous n'attendrions pas du modèle de forêt aléatoire qu'il renvoie une alerte de risque erronée s'il était plus précis, mais il est concevable qu'au niveau individuel, un événement similaire puisse se produire où, dans l'ensemble, le modèle est plus précis, mais fait une prédiction erronée pour certains individus seulement, alors que des modèles démontrés moins précis en feraient une plus juste.

Bien qu'un certain nombre de détails, concernant les variables exactes incluses dans les modèles ou les éléments de conception spécifiques des algorithmes, ne soient pas accessibles au public, soit parce que ces informations sont propriétaires, soit parce qu'elles ne sont pas publiées, nous pouvons néanmoins constater des différences dans les données sous-jacentes et la conception des modèles. Bien que ces considérations sur la transparence rendent difficile l'évaluation de la cohérence de différents outils conçus pour différentes autorités locales au sein de ce système de services décentralisés, nous pourrions nous attendre à ce que ces inégalités puissent avoir des impacts différentiels sur les évaluations des risques et l'assistance qu'elles apportent aux travailleurs. Cela pourrait entraîner des conséquences sur la cohérence de la prise de décision en fonction du lieu de vie. De nombreux détails sur les modèles n'étant pas disponibles – et, lorsqu'ils sont disponibles, avec des différences dans les variables dépendantes et indépendantes, ainsi que dans les ensembles de données sous-jacents –, une comparaison précise n'a pas été possible. Ce qu'indique ce manque de transparence, c'est que si une tentative plus systématique de comparaison de la cohérence des modèles était tentée, au-delà de l'expérience de pensée proposée ici, les limitations des informations disponibles pourraient empêcher cette entreprise. Toutefois, si l'on reprend l'expérience de pensée sur les prêts hypothécaires, cette divergence entre les autorités locales suggère que, toutes choses égales par ailleurs, les évaluations du risque d'un individu peuvent différer entre les autorités locales qui utilisent des modèles différents. Les principaux problèmes sont donc les suivants : il semble que des algorithmes différents soient utilisés dans les différentes autorités locales ; des outils différents dans des autorités locales différentes peuvent prendre des décisions incohérentes entre elles en raison des **(p.288)** différences de modèles ou de données ; et les algorithmes peuvent ne pas être suffisamment transparents pour que ceux qui analysent ces modèles puissent s'en rendre compte.

Les principales conclusions sont les suivantes : si l'on prend en compte que ces algorithmes sont des outils probabilistes, la prise de décision orientée par les algorithmes pourrait différer au niveau individuel en raison des différences dans les

données d'entrée, le modèle ou ses spécifications. Les conséquences sont les suivantes : le choix de l'algorithme peut dépendre d'autres critères que la seule précision ; les cas similaires peuvent ne pas être traités de la même manière d'un algorithme à l'autre ; les mécanismes permettant aux individus de faire appel des décisions orientées par les algorithmes peuvent ne pas être disponibles ; et si le niveau de transparence des algorithmes n'est pas suffisant, il peut être impossible d'établir si une décision est cohérente. Toute entité composée d'autorités locales ou d'une pluralité d'agences pouvant prendre des décisions indépendantes sur l'adoption d'algorithmes pourrait être confrontée à un problème similaire, lié à la cohérence du processus décisionnel.

## E – Discussion

Ces résultats suggèrent que si deux entités utilisent des modèles différents, il est possible qu'un individu fasse l'objet d'une prédiction différente d'une entité à l'autre. Cela est particulièrement pertinent, car les estimations suggèrent qu'une autorité locale sur trois au Royaume-Uni expérimente l'analyse prédictive (Marsh 2019). Cette question est également pertinente d'un point de vue international, car les recherches sur les utilisations de l'IA dans l'Union européenne suggèrent que des conditions similaires pourraient exister entre les États membres (Misuraca and van Noordt 2020; van Noordt and Misuraca 2020). Le Royaume-Uni, et toute juridiction qui espère tirer parti des outils d'apprentissage automatique pour l'évaluation des risques ou l'assistance à la décision, devra réfléchir sérieusement à la mesure dans laquelle il peut tolérer des incohérences en fonction de l'outil particulier choisi dans des contextes où une pluralité d'outils pourrait être adoptée.

Bien que les personnes du secteur comprennent intuitivement que certaines décisions critiques doivent être laissées aux travailleurs sociaux en tant qu'individus, cela s'inscrit dans le contexte de préoccupations selon lesquelles tout outil qui entre dans un processus appartenant au secteur **(p.289)** peut influencer la décision finale que prend un agent (White, Hall, and Peckover 2009). Certains concepteurs d'outils ont indiqué :

« Une machine seule ne doit pas prendre une décision qui a des consé- quences juridiques pour un individu [...] Même en mettant à part l'aspect juridique, je pense qu'il est tout à fait adéquat que l'humain prenne la décision finale parce que [...] il peut y avoir des éléments d'un cas par- ticulier qui sont très propres à ce cas et qui ne sont pas reflétés par le modèle [...] Nous considérons donc beaucoup cela comme une assistance à la décision », et que le discernement de l'agent devrait toujours prendre la main ; il s'agit simplement d'essayer de lui donner un élément d'information supplémentaire pour l'aider à prendre de meilleures décisions [...] » (Bright et al. 2019, 8–9).

Bien que ces outils ne soient pas destinés à remplacer les agents, ils pour- raient influencer les décisions et si les outils sont suffisamment différents, ils pourraient alors suggérer des décisions différentes.

Si les préoccupations relatives à l'équité en matière de discrimination sont légitimes, l'équité en matière de cohérence de la prise de décision est un domaine qui doit être davantage débattu. Dans un système de services sociaux décentralisé, où les organisations indépendantes sont autorisées à expérimenter de nouvelles technologies comme elles le souhaitent, mais où un organisme de réglementation national est chargé d'assurer une certaine cohérence dans la prestation de ces services aux personnes résidant à l'intérieur de son périmètre, la question de l'incohérence des modèles devrait préoccuper les administrateurs de l'État. Les résultats de cette recherche suggèrent, premièrement, qu'une plus grande transparence est nécessaire autour des modèles, de leurs spécifications et des variables qu'ils utilisent et, deuxièmement, que l'État pourrait avoir à jouer un rôle plus proactif dans la définition de normes pour les outils d'apprentissage automatique dans le secteur public, afin que ses citoyens puissent être sûrs que les prédictions, recommandations et décisions qui émanent de ces machines sont cohérentes d'une région à l'autre. Si une plus grande implication du Gouvernement central peut être nécessaire, l'inertie est un risque associé à l'établissement de normes en matière d'apprentissage automatique. Il existe un risque que l'État ne soit pas en mesure de maintenir sa réglementation à jour par rapport à l'évolution de l'innovation et des meilleures pratiques en matière d'apprentissage automatique. Si la cohérence est importante, elle ne doit pas se faire au détriment des outils les plus précis et les plus équitables disponibles, et un équilibre acceptable pour le public devra être trouvé. **(p.290)**

## Conclusion

En conclusion, le développement d'outils d'aide à la décision par l'analyse prédictive dans les collectivités locales britanniques se heurte non seulement aux défis bien documentés liés aux biais (Benthall and Haynes 2019; Eubanks 2017; Kearns et al. 2019; Madras et al. 2019; Obermeyer and Mullainathan 2019; Verma and Rubin 2018), à la transparence (Munro 2019; Pasquale 2015), à l'explicabilité des résultats (Wachter, Mittelstadt, and Russell 2018; Watson 2020), aux choix de conception des développeurs (Janssen and Kuk 2016) et aux compromis en matière d'équité (Lee and Floridi 2021), mais aussi aux défis liés à la cohérence des prédictions algorithmiques d'une collectivité locale à l'autre. Même si les individus comprennent les caractéristiques de conception d'un outil algorithmique ou peuvent recevoir une explication spéculative de la raison pour laquelle une recommandation particulière a été faite dans leur cas, si les outils sous-jacents sont fondamentalement différents d'une autorité locale à l'autre, alors ces individus peuvent connaître des résultats différents ou recevoir des explications différentes et, par conséquent, se demander légitimement s'ils connaissent des trajectoires de service similaires à celles qu'ils connaîtraient s'ils étaient soumis à des modèles dans des autorités locales voisines.

En ce qui concerne les recherches futures, cet article jette les bases d'une recherche empirique quantitative portant sur le comportement technique réel de ces modèles dans des contextes réels afin de déterminer si cette hypothèse sur l'incohérence des modèles est une préoccupation valide. Si les résultats quantitatifs confirment ce problème, des recherches similaires pourraient être menées dans d'autres juridictions

et secteurs, ainsi qu'en ce qui concerne l'impact des algorithmes sur la cohérence de la prise de décision humaine. Il convient également de noter que la présente étude ne porte que sur différents types de modèles. Comme l'indiquent les résultats qualitatifs descriptifs, il existe un certain nombre de façons dont ces modèles peuvent être différents ou peuvent produire des résultats différents. Par exemple, les variables, le réglage intrinsèque au modèle ou les spécifications, les logiciels et les progiciels peuvent tous être légèrement différents entre les instanciations de modèles, ce qui indique que des recherches futures pourraient jeter davantage de lumière sur l'impact de ces facteurs supplémentaires sur l'incohérence. **(p.291)**

Cette recherche suggère que nous devons examiner non seulement l'exactitude, l'équité ou l'explicabilité des résultats des outils, mais aussi la cohérence des résultats entre les différents outils afin de garantir que les individus ne reçoivent pas des prestations différentes en fonction de l'outil qui leur est appliqué. Si l'un des objectifs du service public est de s'assurer que les individus ont le même type d'expérience de service ou obtiennent les mêmes résultats quel que soit le lieu où ils reçoivent ces services, alors une plus grande transparence et des comparaisons des résultats des différents outils algorithmiques sont nécessaires.

**Références bibliographiques**


ACM FAccT Conference. 2018. "FAT* Conference 2018." ACM FAccT Conference. 2018. https://facctconference.org/2018/index.html.

Ada Lovelace Institute. 2020. "Transparency Mechanisms for UK Public-Sector Algorithmic Decision-Making Systems." Nuffield Foundation. https://www.adalovelaceinstitute.org/evidence-review/transparency-mechanisms-for-uk-public-sector-algorithmic-decision-making-systems/.

Allard, Scott W., Emily R. Wiegand, Colleen Schlecht, A. Rupa Datta, Robert M. Goerge, and Elizabeth Weigensberg. 2018. "State Agencies' Use of Administrative Data for Improved Practice: Needs, Challenges, and Opportunities." *Public Administration Review* 78 (2): 240–50. https://doi.org/10.1111/puar.12883.

Alonso, William, and Paul Starr. 1987. *The Politics of Numbers*. Population of the United States in the 1980s. New York: Russell Sage Foundation.

Atkins, Graham. 2020. "Local Government Funding in England." The Institute for Government. March 10, 2020. https://www.instituteforgovernment.org.uk/explainers/local-government-funding-england.

BASW. 2019. "Plan to Trial 'big Data' Analytics in Social Care Sparks Ethics Debate." British Association of Social Workers. February 12, 2019. https://www.basw.co.uk/resources/psw-magazine/psw-online/plan-trial-big-data-analytics-social-care-sparks-ethics-debate.

Benbasat, Izak, David K. Goldstein, and Melissa Mead. 1987. "The Case Research Strategy in Studies of Information Systems." *MIS Quarterly* 11 (3): 369–86. https://doi.org/10.2307/248684.

Benthall, Sebastian, and Bruce D. Haynes. 2019. “Racial Categories in Machine Learning.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 289–98. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287575.

Bigman, Yochanan E., and Kurt Gray. 2018. “People Are Averse to Machines Making Moral Decisions.” *Cognition* 181 (December): 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

Booth, Robert. 2019. “How Bristol Assesses Citizens’ Risk of Harm – Using an Algorithm.” *The Guardian*, October 15, 2019, sec. UK news. https://www.theguardian.com/uk-news/2019/oct/15/bristol-algorithm-assess-citizens-risk-harm-guide-frontline-staff.

Bovens, Mark, and Stavros Zouridis. 2002. “From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control.” *Public Administration Review* 62 (2): 174–84. https://doi.org/10.1111/0033-3352.00168.

Bright, Jonathan, Bharath Ganesh, Cathrine Seidelin, and Thomas Vogl. 2019. “Data Science for Local Government.” Oxford, UK: Oxford Internet Institute, University of Oxford. https://smartcities.oii.ox.ac.uk/data-science-for-local-government-report/.

Celia, Hannah. 2016. “Xantura: Early Help Profiling System.” Xantura - News & Information: Delivering Shared Intelligence, Where and When It Matters Most. February 4, 2016. https://www.xantura.com/points-of-view/early-help-profiling-system.

Centre for Data Ethics and Innovation. 2019. “Submission 18: Response to the Committee on Standards in Public Life’s AI & Public Standards Review.” In *Artificial Intelligence and Public Standards: Written Evidence*. London, UK: Committee on Standards in Public Life. https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-written-evidence.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. “A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:134–48. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/chouldechova18a.html.

Clayton, Vicky. 2019. “Ethics and Machine Learning in Children’s Social Care.” Nesta. *What Works for Children’s Social Care* (blog). February 22, 2019. https://whatworks-csc.org.uk/blog/ethics-and-machine-learning-in-childrens-social-care/.

Clayton, Vicky, Daniel Gibbons, and Michael Sanders. 2019. “Research Protocol: Pilots of Predictive Analytics in Children’s Social Care.” What Works Centre for Children’s Social Care. https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_Predictive_Analytics_Pilots_Protocol_Oct_2019.pdf.

Clayton, Vicky, Michael Sanders, Eva Schoenwald, Lee Surkis, and Daniel Gibbons. 2020. “Machine Learning in Children’s Services: Technical Report.” London, UK:

What Works for Children's Social Care. https://whatworks-csc.org.uk/research-report/machine-learning-in-childrens-services-does-it-work/.

Coston, Amanda, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. "Counterfactual Risk Assessments, Evaluation, and Fairness." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 582–93. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372851.

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. "Risk Assessment and Decision Making in Child Protective Services: Predictive Risk Modeling in Context." *Children and Youth Services Review* 79 (August): 291–98. https://doi.org/10.1016/j.childyouth.2017.06.027.

Dencik, Lina, Arne Hintz, Joanna Redden, and Harry Warne. 2018. "Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services Project Report." Cardiff, UK: Data Justice Lab, Cardiff University. https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf.

Dunleavy, Patrick, and Helen Margetts. 2015. "Design Principles for Essentially Digital Governance." In *111th Annual Meeting of the American Political Science Association, 3 - 6 September 2015, American Political Science Association*. San Francisco, USA: (Unpublished). http://eprints.lse.ac.uk/64125/1/Essentially%20Digital%20Governance.pdf.

Dunleavy, Patrick, Helen Margetts, Simon Bastow, and Jane Tinkler. 2006. *Digital Era Governance: IT Corporations, the State, and e-Government*. Oxford: Oxford University Press.

Durose, Catherine. 2011. "Revisiting Lipsky: Front-Line Work in UK Local Governance." *Political Studies* 59 (4): 978–95. https://doi.org/10.1111/j.1467-9248.2011.00886.x.

Eaton, Lynn. 2019. "Is It Right to Use AI to Identify Children at Risk of Harm?" *The Guardian*, November 18, 2019, sec. Society. https://www.theguardian.com/society/2019/nov/18/child-protection-ai-predict-prevent-risks.

Eaton, Michelle, and Camilla Bertoncin. 2018. "State of Offices of Data Analytics (ODA) in the UK." London, UK: Nesta. https://media.nesta.org.uk/documents/State_of_Offices_of_Data_Analytics_ODA_in_the_UK_WEB_v5.pdf.

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

Financial Times, The editorial board. 2019. "The Squeeze on Local Authority Funding Needs to Be Eased." *Financial Times*, January 31, 2019. https://www.ft.com/content/5801a0d2-23e8-11e9-b329-c7e6ceb5ffdf.

Fluke, John D, Donald J Baumann, Len I Dalgleish, and Homer D Kern. 2014. "Decisions to Protect Children: A Decision Making Ecology." In *Handbook of Child Maltreatment*, edited by Jill E. Korbin and Richard D. Krugman, 463–76. Dordrecht: Springer.

Flyvbjerg, Bent. 2006. "Five Misunderstandings About Case-Study Research." *Qualitative Inquiry* 12 (2): 219–45. https://doi.org/10.1177/1077800405284363.

Foucault, Michel. 1991. “Governmentality.” In *The Foucault Effect: Studies in Governmentality, with Two Lectures by and an Interview with Michel Foucault*, edited by Graham Burchell, Colin Gordon, and Peter Miller. Chicago: University of Chicago Press.

Garrett, Paul Michael. 2005. “Social Work’s ‘Electronic Turn’: Notes on the Deployment of Information and Communication Technologies in Social Work with Children and Families.” *Critical Social Policy* 25 (4): 529–53. https://doi.org/10.1177/0261018305057044.

Gillingham, Philip. 2019. “The Development of Algorithmically Based Decision-Making Systems in Children’s Protective Services: Is Administrative Data Good Enough?” *The British Journal of Social Work*, no. bcz157 (December). https://doi.org/10.1093/bjsw/bcz157.

Goldenfein, Jake. 2019. “Algorithmic Transparency and Decision-Making Accountability: Thoughts for Buying Machine Learning Algorithms.” In *Closer to the Machine: Technical, Social, and Legal Aspects of AI*. Melbourne, Australia: Office of the Victorian Information Commissioner. https://ovic.vic.gov.au/wp-content/uploads/2019/08/closer-to-the-machine-web.pdf.

Graham, Jack. 2017. “London Uses Data to Predict Which Children Will Be Abused.” Apolitical. September 18, 2017. https://apolitical.co/solution_article/london-uses-data-predict-which-children-abuse/.

Green, Ben, and Yiling Chen. 2019. “Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287563.

Hall, Christopher, Nigel Parton, Sue Peckover, and Sue White. 2010. “Child-Centric Information and Communication Technology (ICT) and the Fragmentation of Child Welfare Practice in England.” *Journal of Social Policy* 39 (3): 393–413. https://doi.org/10.1017/S0047279410000012.

Harris, Tom, Louis Hodge, and David Phillips. 2019. “English Local Government Funding: Trends and Challenges in 2019 and Beyond.” London, UK: The Institute for Fiscal Studies. ISBN 978-1-912805-41-9. https://www.ifs.org.uk/publications/14563.

ICO. 2019. “Project ExplAIn Interim Report.” Information Commissioner’s Office. https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/.

Janssen, Marijn, and George Kuk. 2016. “The Challenges and Limits of Big Data Algorithms in Technocratic Governance.” *Government Information Quarterly*, Open and Smart Governments: Strategies, Tools, and Experiences, 33 (3): 371–77. https://doi.org/10.1016/j.giq.2016.08.011.

Karppinen, Kari, and Hallvard Moe. 2012. “What We Talk about When We Talk about Document Analysis.” In *Trends in Communication Policy Research: New Theories, Methods and Subjects*, edited by Manuel Puppis and Natascha Just, 177–93. Bristol, UK: Intellect Books Ltd. https://www.researchgate.net/profile/Kari_Karppinen/publication/299366807_What_We_Talk_about_When_We_Talk_about_Document_Analysis/links/5719cca80

8aed43f63235a46/What-We-Talk-about-When-We-Talk-about-Document-Analysis.pdf.
Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. “An Empirical Study of Rich Subgroup Fairness for Machine Learning.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287592.
Keddell, Emily. 2014. “Current Debates on Variability in Child Welfare Decision-Making: A Selected Literature Review.” *Social Sciences* 3 (4). https://doi.org/10.3390/socsci3040916.
Kim, Gang-Hoon, Silvana Trimi, and Ji-Hyong Chung. 2014. “Big-Data Applications in the Government Sector.” *Commun. ACM* 57 (3): 78–85. https://doi.org/10.1145/2500873.
Kolkman, Daan. 2020. “The (in)Credibility of Algorithmic Models to Non-Experts.” *Information, Communication & Society*, May, 1–17. https://doi.org/10.1080/1369118X.2020.1761860.
Lee, Michelle Seng Ah. 2019. “Context-Conscious Fairness in Using Machine Learning to Make Decisions.” *AI Matters* 5 (2): 23–29. https://doi.org/10.1145/3340470.3340477.
Lee, Michelle Seng Ah, and Luciano Floridi. 2021. “Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-Offs.” *Minds and Machines* 31 (1): 165–91. https://doi.org/10.1007/s11023-020-09529-4.
Leslie, David, Lisa Holmes, Christina Hitrova, and Eleanor Ott. 2020. “Ethics Review of Machine Learning in Children’s Social Care.” London, UK: What Works for Children’s Social Care. https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_Ethics_of_Machine_Learning_in_CSC_Jan2020.pdf.
Lipsky, Michael. 2010. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Updated. New York: Russell Sage Foundation.
Lord Laming, Herbert. 2003. *The Victoria Climbie Inquiry*. Cm (Series) (Great Britain. Parliament). London: Stationery Office.
Lorenz, Lukas, Albert Meijer, and Tino Schuppan. 2021. “The Algocracy as a New Ideal Type for Government Organizations: Predictive Policing in Berlin as an Empirical Case.” *Information Polity* 26 (1): 71–86. https://doi.org/10.3233/IP-200279.
LSCB Haringey. 2009. “Serious Case Review: Baby Peter.” Local Safeguarding Children Board.
Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. “Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 349–58. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287564.
Malomo, Fola, and Vania Sena. 2017. “Data Intelligence for Local Government? Assessing the Benefits and Barriers to Use of Big Data in the Public Sector.” *Policy & Internet* 9 (1): 7–27. https://doi.org/10.1002/poi3.141.
Marsh, Sarah. 2019. “One in Three Councils Using Algorithms to Make Welfare Decisions.” *The Guardian*, October 15, 2019, sec. Society.

https://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits.
Marsh, Sarah, and Niamh McIntyre. 2020. “Nearly Half of Councils in Great Britain Use Algorithms to Help Make Claims Decisions.” *The Guardian*, October 28, 2020, sec. Society. https://www.theguardian.com/society/2020/oct/28/nearly-half-of-councils-in-great-britain-use-algorithms-to-help-make-claims-decisions.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.
McIntyre, Niamh, and David Pegg. 2018. “Councils Use 377,000 People’s Data in Efforts to Predict Child Abuse.” *The Guardian*, September 16, 2018, sec. Society. https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse.
Misuraca, Gianluca, and Colin van Noordt. 2020. “AI Watch - Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU.” EUR 30255 EN. JRC Science for Policy Report. Luxembourg: Publications Office of the European Union. doi:10.2760/039619.
Munro, Eileen. 2019. “Predictive Analytics in Child Protection.” *CHESS Working Paper*, Produced as part of the Knowledge for Use (K4U) Research Project, No. 2019-03 (April). https://www.researchgate.net/publication/332528200_Predictive_analytics_in_child_protection.
Noordt, Colin van, and Gianluca Misuraca. 2020. “Exploratory Insights on Artificial Intelligence for Government in Europe.” *Social Science Computer Review*, December, 0894439320980449. https://doi.org/10.1177/0894439320980449.
Obermeyer, Ziad, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm That Guides Health Decisions for 70 Million People.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 89. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287593.
Orlikowski, Wanda J., and Jack J. Baroudi. 1991. “Studying Information Technology in Organizations: Research Approaches and Assumptions.” *Information Systems Research* 2 (1): 1–28. https://doi.org/10.1287/isre.2.1.1.
Pasquale, Frank. 2015. *The Black Box Society*. Harvard University Press.
Peckover, Sue, Christopher Hall, and Sue White. 2009. “From Policy to Practice: The Implementation and Negotiation of Technologies in Everyday Child Welfare.” *Children & Society* 23 (2): 136–48. https://doi.org/10.1111/j.1099-0860.2008.00143.x.
Pénicaud, Soizic. 2021. “Building Public Algorithm Registers: Lessons Learned from the French Approach.” Open Government Partnership. *Open Algorithms Blog Series* (blog). May 12, 2021. https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/.
Pentland, Alex. 2014. *Social Physics: How Good Ideas Spread: The Lessons from a New Science*. New York: Penguin Press.
Pollitt, Christopher. 2009. “Bureaucracies Remember, Post-Bureaucratic Organizations Forget?” *Public Administration* 87 (2): 198–218.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. “The Algorithmic Automation Problem: Prediction, Triage, and Human Effort.” *ArXiv:1903.12220 [Cs]*, March. http://arxiv.org/abs/1903.12220.

Rawls, John. 2009. *A Theory of Justice*. Harvard university press.

Redden, Joanna, Lina Dencik, and Harry Warne. 2020. “Datafied Child Welfare Services: Unpacking Politics, Economics and Power.” *Policy Studies* 41 (5): 507–26. https://doi.org/10.1080/01442872.2020.1724928.

Sanders, Michael, James Lawrence, Daniel Gibbons, and Paul Calcraft. 2017. “Using Data Science in Policy.” London: Behavioural Insights Team. https://www.bi.team/wp-content/uploads/2017/12/BIT_DATA-SCIENCE_WEB-READY.pdf.

Scott, James C. 1998. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale Agrarian Studies. New Haven: Yale University Press.

Sena, Vania, Anna Crispe, Seb Smith, Olga Sergushova, and Andrei Toderas. 2019. “Application of Predictive Analytics in Social Care: Tools to Support Decision Making.” Conference Presentation presented at the Catalyst Conference, University of Essex, June 26. https://www.essex.ac.uk/-/media/documents/research/catalyst/catalyst-workshope-analytics-social-care.pdf.

Shafiq, Wajid. 2018. “Xantura: Media Coverage - a Welcome Debate / Discussion.” Xantura - News & Information: Delivering Shared Intelligence, Where and When It Matters Most. September 19, 2018. https://www.xantura.com/points-of-view/media-coverage-welcome-debate-discussion.

———. 2020. “Machine Learning Can Deliver Better Outcomes for Children and Families.” *Community Care* (blog). September 21, 2020. https://www.communitycare.co.uk/2020/09/21/data-sharing-supported-machine-learning-can-deliver-better-outcomes-children-families/.

Sheridan, Ed. 2018. “Hackney Council Pays £360k to Data Firm Whose Software Profiles Troubled Families.” *Hackney Citizen*, October 18, 2018. https://www.hackneycitizen.co.uk/2018/10/18/council-360k-xantura-software-profiles-troubled-families/.

Snow, Thea. 2021. “From Satisficing to Artificing: The Evolution of Administrative Decision-Making in the Age of the Algorithm.” *Data & Policy* 3: e3. https://doi.org/10.1017/dap.2020.25.

SocialWhatNow. 2019. “Predictive Analytics and the What Works Centre for Children’s Social Care — Connecting Some Dots….” *Medium* (blog). October 17, 2019. https://medium.com/@SocialWhatNow/predictive-analytics-and-the-what-works-centre-for-childrens-social-care-connecting-dots-the-old-6fd7d094771d.

Stevenson, Luke. 2018. “Artificial Intelligence: How a Council Seeks to Predict Support Needs for Children and Families.” *Community Care* (blog). March 1, 2018. https://www.communitycare.co.uk/2018/03/01/artificial-intelligence-council-seeks-predict-support-needs-children-families/.

Symons, Tom. 2016a. “Datavores of Local Government: Using Data to Make Services More Personalised, Effective and Efficient - Discussion Paper.” London, UK: Nesta. https://www.nesta.org.uk/report/datavores-of-local-government/.

———. 2016b. “Wise Council: Insights from the Cutting Edge of Data-Driven Local Government.” London: Nesta. https://www.nesta.org.uk/report/wise-council-insights-from-the-cutting-edge-of-data-driven-local-government/.

Thorne, Sally, Louise Jensen, Margaret H. Kearney, George Noblit, and Margarete Sandelowski. 2004. “Qualitative Metasynthesis: Reflections on Methodological Orientation and Ideological Agenda.” *Qualitative Health Research* 14 (10): 1342–65. https://doi.org/10.1177/1049732304269888.

Timulak, Ladislav. 2014. “Chapter 33 | Qualitative Meta-Analysis.” In *The SAGE Handbook of Qualitative Data Analysis*, edited by Uwe Flick, 481–95. London: SAGE Publications Ltd. https://doi.org/10.4135/9781446282243.

Toderas, Andrei, and Mina Manning. 2017. “The Benefits of Predictive Analytics in Councils.” University of Essex: Catalyst. https://www.essex.ac.uk/-/media/documents/research/catalyst/the-future-of-predictive-analytics-in-councils.pdf.

Tsoukas, Haridimos. 2009. “Craving for Generality and Small-N Studies: A Wittgensteinian Approach towards the Epistemology of the Particular in Organization and Management Studies.” In *The SAGE Handbook of Organizational Research Methods*, edited by David A Buchanan and Alan Bryman. Organizational Research Methods. Los Angeles: SAGE.

Turner, Alex. 2019a. “County to Trial Predictive Algorithms in Children’s Social Work.” *Community Care* (blog). June 14, 2019. https://www.communitycare.co.uk/2019/06/14/county-becomes-latest-authority-trial-predictive-algorithms-childrens-social-work/.

———. 2019b. “Algorithms in Children’s Social Care: Call for Understanding of Risks and Benefits.” *Community Care* (blog). November 15, 2019. https://www.communitycare.co.uk/2019/11/15/using-algorithms-childrens-social-care-experts-call-better-understanding-risks-benefits/.

University of Essex. 2019. “Catalyst Project - Improving Community Services for Vulnerable People.” University of Essex. 2019. https://www.essex.ac.uk/research/showcase/catalyst-project.

Verma, Sahil, and Julia Rubin. 2018. “Fairness Definitions Explained.” In *Proceedings of the International Workshop on Software Fairness*, 1–7. FairWare ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3194770.3194776.

Vogl, Thomas M. 2021. “Artificial Intelligence in Local Government: Enabling Artificial Intelligence for Good Governance in UK Local Authorities.” 2. Oxford, UK: Oxford Commission on AI & Good Governance. https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2021/04/Local-Government-report.pdf.

Vogl, Thomas M., Cathrine Seidelin, Bharath Ganesh, and Jonathan Bright. 2020. “Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities.” *Public Administration Review* 80 (6): 946–61. https://doi.org/10.1111/puar.13286.

Wachter, Sandra, Brent Daniel Mittelstadt, and Chris Russell. 2018. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR.” *Harvard Journal of Law & Technology (Harvard JOLT)* 31 (2): 841–88.

Watson, David. 2020. “Conceptual Challenges for Interpretable Machine Learning.” SSRN Scholarly Paper ID 3668444. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3668444.

Weber, Max. 1968. *Economy and Society: An Outline of Interpretive Sociology*. Edited by Guenther Roth and Claus Wittich. New York: Bedminster Press.

White, Sue, Chris Hall, and Sue Peckover. 2009. “The Descriptive Tyranny of the Common Assessment Framework: Technologies of Categorization and Professional Practice in Child Welfare.” *The British Journal of Social Work* 39 (7): 1197–1217. https://doi.org/10.1093/bjsw/bcn053.

WW for CSC. 2018. “What Works for Children’s Social Care.” What Works for Children’s Social Care. 2018. https://whatworks-csc.org.uk/.

———. 2019. “Call for Partners to Pilot Potential of Analysing Case Notes to Assist Social Workers.” *What Works for Children’s Social Care* (blog). February 7, 2019. https://whatworks-csc.org.uk/blog/call-for-partners-to-pilot-potential-of-analysing-case-notes-to-assist-social-workers/.

Xantura. 2016. “Xantura: Children’s Safeguarding.” Xantura - News & Information: Delivering Shared Intelligence, Where and When It Matters Most. 2016. https://www.xantura.com/focus-areas/childrens-safeguarding.

———. 2018. “Xantura: Maximising the Impact of Early Help Resources.” Xantura - News & Information: Delivering Shared Intelligence, Where and When It Matters Most. September 24, 2018. https://www.xantura.com/points-of-view/maximising-imact-early-healp-resources.

Yang, Tung-Mou, and Terrence A. Maxwell. 2011. “Information-Sharing in Public Organizations: A Literature Review of Interpersonal, Intra-Organizational and Inter-Organizational Success Factors.” *Government Information Quarterly* 28 (2): 164–75. https://doi.org/10.1016/j.giq.2010.06.008.

Zacka, Bernardo. 2017. *When the State Meets the Street: Public Service and Moral Agency*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy & Technology* 32 (4): 661–83. https://doi.org/10.1007/s13347-018-0330-6.