Privacy and fairness issues with algorithmic decision systems

Louis Béziaud

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

UQÀM

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES
EN COTUTELLE AVEC
L'UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Informatique*

Par
Louis BÉZIAUD

Privacy and fairness issues with algorithmic decision systems

Thèse présentée et soutenue à Rennes, le 6 décembre 2023
Unité de recherche : IRISA (UMR CNRS N° 6074)

Rapporteurs avant soutenance :

Alexis TSOUKIÀS — Directeur de recherche – CNRS
Benjamin NGUYEN — Professeur – INSA Centre Val de Loire

Composition du Jury :

Président :	Benjamin NGUYEN	
Examinateur·trice·s :	Alexis TSOUKIÀS	Directeur de recherche – CNRS
	Benjamin NGUYEN	Professeur – INSA Centre Val de Loire
	Luis GALÁRRAGA	Chercheur – INRIA Rennes
	Sihem AMER-YAHIA	Directrice de recherche – CNRS
Co-dir. de thèse :	Tristan ALLARD	Maître de conférences – Université de Rennes
Co-dir. de thèse :	Sébastien GAMBS	Professeur – Université du Québec à Montréal

Table of contents

This dissertation begins with a summary in French. The rest of this document is written in English, and starts at page 13.

Résumé en français

Introduction

La multiplication des collectes de données personnelles, que ce soit lors de transactions commerciales, de déplacements, d'interactions avec les institutions ou de communications, est accompagnée d'un recours tout aussi massif à des systèmes algorithmiques. En particulier, ces systèmes algorithmiques sont de plus en plus utilisés pour prendre des décisions impactant les individus, les groupes et la société dans son ensemble. Ainsi, ils prennent une place de plus en plus prépondérante dans de nombreux domaines comme outil d'assistance à la prise de décision ou comme agent décisionnel à part entière, que ce soit pour évaluer les curriculums vitæ de candidat·es à Amazon [Das18], pour apparier étudiant·es et établissements avec Parcoursup [FPZ19], pour évaluer le risque de récidive des accusé·es comme le fait le logiciel COMPAS [BD18] utilisé aux États-Unis ou pour l'attribution d'organes par Eurotransplant [Sch+18] en Europe. Cette ubiquité soulève d'importantes préoccupations éthiques et sociales, notamment concernant la protection de la vie privée et l'équité de ces approches, comme le montrent les nombreuses réglementations établies comme le RGPD [Eur16] ou en préparation comme l'AI Act [Eur21], ainsi que les multiples ouvrages abordant ces questions [Zho+20 ; Bir+22].

La particularité sociale de ces enjeux, à la fois par leur contexte d'application et par la nature des questions, rend leur étude sous l'angle technique parcellaire. Si des approches comme l'évaluation des risques permettent de quantifier dans une certaine mesure l'impact tant pour la vie privée que la discrimination, celle-ci ne peut se faire qu'en connaissance des dangers. Cependant, les définitions de ces risques évoluent à travers le temps, les cultures et les positions politiques. Les étudier nécessite de pouvoir modéliser des systèmes sociotechniques complexes et cette modélisation peut être elle-même porteuse de biais de par le choix de consider ou non certains éléments [Ave11]. Par ailleurs, les dimensions d'interprétation et de contexte de concepts comme la "vie privée" et la "non discrimination" ne peuvent pas être retranscrites facilement de façon formelle, laissant place à des redéfinitions algorithmiques comme discuté par des travaux récents [Och19 ; Hof19 ; Kul+20]. Des conférences comme ACM FAccT [ACM18] tentent de répondre à cet enjeu en faisant se rejoindre les expertises et outils de plusieurs communautés, avec de nombreux articles mélant des auteur·ices issues des domaines de l'informatique, de la linguistique, du droit ou encore de la sociologie. Cette première réponse nous semble cependant souffrir de sa limitation au champ technique, amenant certes la dimension sociale en perspective, mais se limitant à des réponses techniques. On peut prendre par exemple le cas du cadre analytique de l'intersectionnalité [Cre89], identifié par l'universitaire afroféministe américaine Kimberlé Williams Crenshaw, qui permet de

considérer les différentes formes de domination, d'oppression et de discrimination non pas comme cloisonnées, mais comme un système interconnecté, avec l'idée que celles-ci sont "plus grandes que la somme de leurs parties" [PB20]. Si cette notion est décrite et étudiée en longueur par de multiples ouvrages en sciences humaines et sociales, elle est formalisée dans plusieurs articles techniques [Fou+20 ; Mor+19 ; Kan+22 ; ML22] comme la *simple* considération d'une multiplicité d'attributs sensibles, par exemple l'intersection de genre et de race. Bien que ces propositions de formulations permettent un premier pas vers l'utilisation d'un concept abstrait dans un environnent technique, la réduction en inéquations, par exemple formulée sur le modèle de la confidentialité différentielle et appliquée à des sous-groupes de la population dans [Fou+20], ne peut conserver le contexte et les principes méthodologiques de l'approche. La complexité d'utilisation est illustrée par [Bau+21] qui compare l'utilisation du cadre intersectionnel par plusieurs études quantitatives, ou par [VK22] qui propose une première définition empirique. Les considérations sociales fortes rendent particulièrement crucial d'évaluer ce qui est perdu lors de la technisisation des concepts, au delà de l'exemple de l'intersectionnalité utilisé ici.

Cette thèse cherche à étudier la protection de la vie privée et l'équité sous l'angle de leur utilisation pratique, mais aussi de leurs implications sociétales. S'il est difficile de parler de technique sans être technicien·e et tout autant de parler d'éthique sans être éthicien·ne, on gardera ici une approche d'informaticien, en ayant pour objectif d'aller vers une ouverture sur les sciences sociales.

Protection de la vie privée

En matière de protection de la vie privée, les premières réponses se sont faites avec des mécanismes empiriques. Par exemple, la k-anonymité introduite en 2002 [Swe02] cache les profils individuels en les regroupant en grappes d'au moins k personnes. Ce modèle de vie privée a été successivement raffiné avec des volontés de contrôler l'utilité (plus précisément la perte d'information) et de répondre à des attaques, amenant à la l-diversité [Mac+07] puis à la t-proximité [LLV07].

En parallèle, la confidentialité différentielle [Dwo06] a été proposée en 2006 avec un objectif formel devant représenter le concept de vie privée auquel obéit la définition. Plus précisément, un mécanisme respecte la vie privée (au sens de la confidentialité différentielle) si la présence ou l'absence d'un individu dans la base de données fournie en entrée au mécanisme impacte le calcul de manière indiscernable, selon un budget alloué. En d'autres mots, étant donné le résultat d'une requête, un·e attaquant·e ne gagne qu'une information limitée permettant d'inférer la présence ou absence de l'individu.

Si cette définition est considérée, dans ce travail comme dans beaucoup d'autres, comme la référence, de très nombreuses variations ont été proposées [DP20] afin de répondre à différents modèles d'adversaire, de connaissances auxiliaires ou d'autres structures de données. À la question du choix d'une définition s'ajoute celle du choix d'un budget (et des variables additionnelles d'autres modèles), qui définit la quantité d'information pouvant fuiter, avec des tentatives de recommandation académiques [Hsu+14 ; LC11] et des utilisations pratiques différentes dans l'industrie [Tan+17].

Pour résumer, la vie privée dans le domaine de l'informatique vise à y formaliser des desiderata sociaux en des objectifs numériques (formels), qui génèrent ensuite leurs propres questions techniques. On transforme ainsi des concepts parfois élusifs de respect de la vie privée et d'équité, issues d'attentes individuelles, de la morale et du droit, en des propriétés mathématiques des algorithmes et de leurs sorties [Och20].

Discrimination

Du côté de la non-discrimination, aussi appelé parfois équité, il s'agit d'un enjeu étudié plus récemment en informatique que le domaine de la vie privée, mais on retrouve des similarités dans l'approche suivie. Ainsi, les nombreuses façons de conceptualiser la discrimination et sa résolution se sont matérialisées dans de multiples redéfinitions algorithmiques de notions complexes. Les définitions les plus communes sont celles dites "de groupe", qui prennent majoritairement comme cadre les décisions issues d'algorithmes d'apprentissage supervisé, en cherchant à comparer la sortie d'un système (c'est-à-dire ses décisions) à des données de référence. Dans ce cadre, la discrimination de groupe se définit comme l'écart entre un certain indicateur, comme le taux de décisions positives ou le taux d'erreur, évalué entre deux groupes. De nombreuses définitions appartenant à cette famille existent [Meh+22] et sont généralement mutuellement exclusives [FSV16]. Pour résumer, la question est de démontrer que le système n'ajoute pas de discrimination, et non pas qu'il n'est pas discriminatoire à l'origine.

Une autre conception propose une vision "individuelle" [Dwo+12] de non-discrimination, en exigeant que deux profils "proches" reçoivent des décisions similaires, étant donnée une notion de similarité entre les individus. La définition de non-discrimination repose alors sur cette mesure de similarité qui doit être elle-même non-discriminante, sans qu'elle puisse être définie à l'aide de ce modèle, posant naturellement un problème d'applicabilité. Il est intéressant de noter que Cynthia Dwork, première autrice de ce travail, est aussi à l'origine de la définition de confidentialité différentielle. On peut aussi citer les définitions utilisant un modèle causal, l'objectif étant alors de supprimer l'impact des variables confondantes dans la prise de décision, c'est-à-dire des variables qui influencent à la fois la décision et les variables d'entrée.

Cette multiplicité de définitions et de concepts est similaire dans une certaine mesure à ce qu'il est possible d'observer pour la vie privée, à la différence qu'aucune définition n'émerge comme préférable techniquement ou socialement comme le fait la confidentialité différentielle en protection de la vie privée. En particulier, il n'est pas désirable ou possible d'aboutir à une telle convergence, à la fois de par les incompatibilités formelles [FSV16] de concepts "désirables", ainsi que par la dépendance entre l'intérêt pour un concept et l'agent qui le formule [Nar18], laissant de fait place à des biais dans le choix de la définition elle-même.

Nous faisons donc le constat de la difficulté à définir, choisir et utiliser des définitions de vie privée et de discrimination dans des contextes sociotechniques. Ne cherchant pas, ni ne pensant qu'il soit possible ou désirable de trouver, de réponse technique à ces enjeux, on s'attache dans cette thèse à étudier les limites et à présenter des solutions dont l'utilisation pourrait se faire en dehors du cercle technique, dans une perspective

d'y lier les personnes sociales et les outils techniques.

Compromis entre vie privée et transparence

Dans une première contribution, nous examinons le conflit entre la vie privée et la transparence lors de la publication de procédures judiciaires. Plus précisément, nous proposons de considérer la transparence comme un desideratum humain d'une part et la vie privée comme une exigence pour le traitement automatisé d'autre part. Les décisions de justice sont aujourd'hui publiées massivement en ligne, à la fois dans un objectif de transparence et d'accessibilité de la justice, et pour répondre à la demande d'accès à des données nécessaire au développement d'outils automatisés pour le droit appelés "legal techs". Cet objectif de publication se heurte aux obligations de respect de la vie privée, pour lesquels nous identifions deux axes de tensions : individuel et massif.

La protection "individuelle" des jugements correspond à l'approche adoptée aujourd'hui de caviardage d'informations qui permette de façon évidente d'identifier les participant·e·s d'un jugement selon les règles qui encadrent la protection de chacun·e. Cette transformation peut être appliquée directement aux jugements en langage naturel et permet une grande utilité en conservant le sens à l'exception de quelques termes comme les noms qui peuvent être remplacés par exemple par des initiales. Cependant, si cette approche empirique rend potentiellement moins évidente une réidentification, elle n'apporte aucune garantie. Nous présentons plusieurs extraits de jugements qui permettent de mettre en valeur ces lacunes, ainsi que les difficultés techniques qui se présenteraient pour y répondre. Une technique comme la confidentialité différentielle permet de répondre à ce problème, mais n'est pas applicable directement au langage.

Concernant la protection "massive", nous partons du constat que le traitement massif de documents correspond au traitement automatique, et que celui-ci se fait en grande majorité après un traitement visant à structurer les données, par exemple en fréquence de mots. Cette séparation des pratiques nous permet de proposer une séparation des besoins en matière de vie privée, et de suggérer que des techniques de protection adaptées au traitement massif comme la confidentialité différentielle soient utilisées pour ce qui est de la publication massive, tout en conservant des méthodes limitées, mais répondant à l'exigence d'utilité pour ce qui est de l'accès individuel aux jugements. Nous proposons finalement une description d'architecture qui permettrait de mettre en place un tel système et identifions les composants particuliers qui seraient nécessaires à sa conception dans un cadre réel.

Évaluation empirique de la protection de la vie privée

Dans une deuxième contribution, nous proposons un cadre pour organiser des compétitions d'attaque de mécanismes de publication de données respectueux de la vie privée afin de mieux établir leurs comportements et limites dans la pratique. Les compétitions sont couramment utilisées dans la communauté de l'apprentissage automatique pour

stimuler le développement de solutions pratiques et efficaces à des problèmes difficiles ou nouveaux, ainsi que dans la communauté de la sécurité à des fins de formation ou pour l'évaluation de sécurité des infrastructures existantes. Une telle tradition n'existe par contre pas encore dans le domaine de la protection de la vie privée. Toutefois, ces dernières années, plusieurs défis axés sur les mécanismes de publication respectueux de la vie privée ont été organisés.

Certains se sont principalement concentrés sur l'aspect de la défense, attendant des algorithmes respectant des garanties formelles telles que la confidentialité différentielle tout en atteignant des niveaux d'utilité élevés. D'autres considèrent de plus une phase d'attaques sur les ensembles de données assainis générés par les participants. Dans ce cas, la première phase est ainsi consacrée à la conception d'algorithmes d'assainissement, tandis que la seconde consiste à attaquer les données assainies à l'aide des algorithmes développés au cours de la première phase.

Si les phases d'assainissement ont permis de comprendre les compromis entre protection de la vie privée et utilité, et ont conduit à la mise en œuvre des algorithmes d'assainissement de pointe ou même de nouveaux algorithmes, les résultats des phases d'attaque ont généralement été plus mitigés. Par exemple, les organisateurs du défi "Hide-and-Seek" ont rapporté des résultats d'attaques équivalents à des réponses aléatoires [Jor+20]. Nous pensons que l'une des raisons de cette situation est que pour réussir, la phase d'attaque doit permettre aux participants de consacrer suffisamment de temps à la conception, à la mise en œuvre et à l'évaluation de l'attaque.

Pour ce faire, nous proposons de mettre en place une série de défis en matière de publication de données respectueuse de la vie privée, spécifiquement adaptés aux attaques contre les algorithmes d'assainissement [ABG23c]. En particulier, nous souhaitons concentrer l'énergie et l'expertise des participants contre un petit ensemble d'algorithmes d'assainissement de pointe soigneusement choisis en donnant suffisamment de temps aux participants pour concevoir et tester soigneusement leurs attaques. Finalement, nous souhaitons aussi permettre l'exploration d'un large éventail de modèles d'adversaires possibles et de connaissances de base. Cette proposition se veut complémentaire des compétitions qui se concentrent sur la partie assainissement, permettant ainsi d'aboutir à un pipeline complet de conception et d'évaluation défense-attaque.

Pour permettre un système complet et générique adaptable aux différents cadres d'attaques, nous définissons un ensemble d'abstractions sous la forme d'un graphe de règles de transformation des données d'entrée et soumissions des participant·es vers leur évaluation. La première édition du défi [ABG23b], qui sert d'implémentation de référence, se concentre sur l'inférence d'appartenance [Hu+22 ; SOT22] dans les données tabulaires générées synthétiquement tout en satisfaisant la confidentialité différentielle [DR14].

Simuler l'impact de la discrimination à long terme

Dans une troisième contribution, nous nous concentrons sur les limites des définitions techniques d'équité et utilisons une simulation ancrée dans la réalité comme moyen d'observer leur impact à long terme sur l'ensemble d'un système. Les définitions de dis-

crimination utilisées aujourd'hui proposent souvent une conception "instantanée" dans laquelle la sortie de l'algorithme est considérée sans prendre en compte son futur, à savoir le contexte dans lequel la prédiction pourra être utilisée pour prendre d'autres décisions. Notre premier pas vers la considération de l'impact d'une décision et de contraintes de non-discrimination sur les individus, les groupes et le système dans son ensemble se fait par une simulation. En effet, cet outil technique permet de répondre, bien que de façon imparfaite, à la nécessité de pouvoir observer le comportement de mécanismes de non-discrimination sans pour autant nécessiter d'expérimentation immorale sur des individus ou impossible sur une société.

Plus précisément, nous adaptons un modèle proposé par Reardon et al. [Rea+18] et issu du domaine de l'éducation dont les valeurs sont adaptées du système américain et dans lequel des étudiant·es de différents groupes sensibles, compétences et revenus postulent et sont admis ou non dans des écoles de qualité variable. Ce modèle prend en compte l'évolution des écoles d'une itération à l'autre en fonction de la qualité de ses étudiant·es. En revanche, le modèle ne considère pas l'impact des décisions sur les étudiant·es. Nous intégrons ce mécanisme pour pouvoir ainsi prendre en compte la dimension temporelle. La simulation finale comprend deux mécanismes de dépendances temporelles : la boucle de renforcement des systèmes de prise de décision et l'effet "boule de neige" pour les individus. La première est similaire à sa définition dans un contexte d'apprentissage par renforcement, à savoir que les décisions produites par un système ont un effet sur l'environnent, qui lui-même impacte l'état du futur système. L'effet boule de neige consiste à considérer une accumulation de décisions du point de vue des individus : une décision reçue est potentiellement utilisée ensuite pour prendre une autre décision, comme c'est le cas par exemple pour le passage d'un lycée à une université. À partir de cette simulation reproduite et modifiée, nous souhaitons observer le comportement de mécanismes simples de non-discrimination et comment leur effet et mesuré ou nous par différentes métriques.

Contributions

Afin de mettre en évidence les questions fondamentales soulevées par la technicisation des défis sociaux et comme tentative de ramener ces défis sociotechniques dans leur contexte social, cette thèse apporte donc les contributions suivantes :

- [ABG20a ; ABG20b] étudient les difficultés en matière de respect de la vie privée rencontrées lors de la publication des décisions de justice, illustrant l'inadéquation de méthodes génériques dans certains contextes ;
- [ABG23c] fait la description et la démonstration d'un cadriciel visant à faciliter la mise en place de compétitions d'attaque de mécanismes de protection de la vie privée pour mieux en cerner les limites. [ABG23b] est une compétition internationale d'attaque de mécanismes de génération de données synthétiques tabulaires respectueux de la vie privée, organisée sur cette base à l'été 2023 ;

- [ABG23a] décrit notre effort de reproduction d'une simulation que nous adaptons par la suite pour étudier les impacts à long terme des mécanismes discriminatoires.

Une liste complète est présentée en page 11.

List of contributions

Authors are listed in alphabetical order, with the exception of [AGB21] for which the authorship sequence was decided by the editors. I am the main author of [ABG20a; ABG20b; ABG23a; ABG23b; ABG23c].

Peer-reviewed international journals

[ABG23a] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "[~Re] Simulating Socioeconomic-Based Affirmative Action". In: *ReScience C* 9.1 (2023). Ed. by Olivia Guest. DOI: 10.5281/zenodo.10255346. HAL: hal-04328511.

Peer-reviewed international workshops and demonstrations

[ABG20a] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "Online publication of court records: circumventing the privacy-transparency trade-off". In: *CoRR* (2020). Presented at ICML 2020's Law and Machine Learning Workshop. arXiv: 2007.01688.

[ABG20b] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "Publication of Court Records: Circumventing the Privacy-Transparency Trade-Off". In: *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers*. Ed. by Víctor Rodríguez-Doncel et al. Vol. 13048. Lecture Notes in Computer Science. Springer, 2020, pp. 298–312. DOI: 10.1007/978-3-030-89811-3_21. HAL: hal-03225201.

[ABG23c] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "SNAKE Challenge: Sanitization Algorithms under Attack". In: *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management*. Birmingham, United Kingdom: Association for Computing Machinery, Oct. 21, 2023. DOI: 10.1145/3583780.3614754. HAL: hal-04228115.

Book chapters

[All+20] Tristan Allard et al. "Ouvrir La Boîte Noire Des Algorithmes de Personnalisation". In: *Le Profilage En Ligne : Entre Libéralisme et Régulation*. Ed. by Alexandra Bensamoun, Maryline Boizard, and Sandrine Turgis. Libre Droit. Mare & Martin, Oct. 15, 2020, pp. 211–231. ISBN: 978-2-84934-466-8.

Dissemination

[AGB21] Tristan Allard, Sébastien Gambs, and Louis Béziaud. "La Confidentialité Différentielle, Garante de l'anonymat". In: *Pour la Science*. Hors-Série 112 (July 8, 2021).

Organization of international challenges

[ABG23b] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "1st Edition of the SNAKE Challenge. SNAKE #1: SaNitization Algorithm under attacK ϵ". collocated with the 13ième Atelier sur la Protection de la Vie Privée (APVP'23). 2023.

1. Introduction

Algorithms in society

The proliferation of personal data collection, whether for commercial transactions, travel, interactions with institutions or communications, is accompanied by an equally massive use of algorithmic systems. In particular, these algorithmic systems are increasingly used to perform decisions impacting individuals, groups, and the society as a whole. As a result, they have taken an increasingly prominent role in many fields as a decision-making tool, or as a decision-making "agent" in their own right. For example, they are used to assess the curricula vitae of candidates for Amazon [Das18], to match students and institutions with Parcoursup [FPZ19], to assess the risk of recidivism of accused persons as done by the COMPAS [BD18] software used in the USA, or for organ allocation by Eurotransplant [Sch+18] in Europe. This ubiquity raises important ethical and social concerns, especially regarding the privacy and fairness of these approaches, as shown by the many regulations established such as the RGPD [Eur16] or in preparation such as the AI Act [Eur21], as well as the multiple works addressing these issues [Zho+20; Bir+22].

Sociotechnical issues

The social particularity of these issues, both in terms of their context of application and the nature of the questions, tends to make them difficult to study from a technical angle. While approaches such as risk assessment make it possible to quantify to some extent the impact on both privacy and discrimination, this can only be done in full awareness of the dangers. However, definitions of these risks evolve over time, across cultures, and political positions. Their study requires the ability to model complex sociotechnical systems, and this modeling can itself be biased by the choice of whether or not to consider certain elements [Ave11]. Furthermore, the interpretive and contextual dimensions of concepts such as "privacy" and "fairness" cannot be easily transcribed formally, leaving room for algorithmic redefinitions as discussed by recent works [Kul+20; Och19; Hof19], which is to be understood as an unfaithful translation of the concept of discrimination that was originally presented. Conferences such as ACM FAccT [ACM18] attempt to meet this challenge by bringing together the expertise and approaches of several communities, with many papers mixing authors from the fields of computer science, linguistics, law, ethics, and sociology. Nevertheless, we believe that this first response suffers from its limitation to the technical field, bringing the social dimension into perspective, but limiting itself to technical answers. Social considerations make

it particularly crucial to assess what is lost when concepts are technicized, beyond the example of intersectionality used here. This thesis seeks to study privacy and fairness from the angle of their practical use, but also of their societal implications. Since it is difficult to talk about technology without being a technician, and just as difficult to talk about ethics without being an ethicist, the approach adopted here is that of a computer scientist, with the aim of opening up to the social sciences. As far as computer science is concerned, the main aim is to formalize social desiderata into (formal) numerical objectives, which then generate their own technical questions. This transforms elusive definitions of privacy and fairness, derived from individual expectations, morality and law, into mathematical properties of algorithms and their outputs [Och20].

This point is the underlying theme of this thesis, by which we want to highlight the importance of the social context when considering these technical issues, thus requiring sociotechnical answers.

Decision systems

Interactions between individuals are governed by decisions, with decisions originating from institutions or individuals, applied to one self or another, applied directly to group of individuals or indirectly through its members. In addition, decisions taken in the past impact future decisions [Nas90]. Institutions change over time with history, norms, or jurisprudence for the example of the judicial system. Individuals evolve by themselves [Ara13] and through interactions with both institutions or groups of individuals, which themselves are thought over time with concepts such as social reproduction through culture and education [BP64; BP70], or economic factors [Tom21].

Computers have naturally become a part of decision processes [Rou+22]. This can occur indirectly, when computers process and create information used to take decisions (e.g., producing statistics for weather forecasts) or as decision support system producing recommendations that can be used to take the final decision (such as the COMPAS software [BD18] used by judges to assess recidivism risk), or directly, which corresponds to the situation in which an algorithm is used to take the full decision which is then applied as-is. Note that the algorithm is rarely applying decisions that impact individuals, but rather it produces predictions, which are then used to make decisions by a human or by an automatic system. However, the frontier between direct and indirect decision-making is blurry, if for example judges blindly follow the COMPAS assessment, without ever refusing its prediction [Sim18].

The first algorithmic decision systems were called *expert systems*. They can be defined as a translation of the set of rules followed by experts of some domain to produce a decision. A typical representation would be a set of rules evaluated by an inference engine, using a knowledge base constructed by domain experts. For example, JUDGE [Mit+86] is a rule-based legal expert system introduced in 1986 that deals with sentencing in the criminal legal domain for offenses relating to murder, assault and manslaughter. SHYSTER [Tyr00] is another such tool introduced in 1996, highlighting by itself the issue of such tool: (algorithmic) pragmatism as a way to be unbiased. Since then, *machine*

learning has progressively replaced expert systems, with the main difference being the reliance on data to extract a set of decision patterns, rather than requiring the process of experts to be encoded. Thus, the aim is to use previous existing decisions directly to infer the rules, skipping the knowledge base. The quality of the production will be impacted both by the choice of learning algorithm and parameters as well as by the quality of training dataset.

In decision-making, high-stakes decisions are described by the existence of a possible large loss and a high cost of reversing the decision–when even possible [Rud18]. For instance, high-stakes decisions can be found in the context of justice–such as a jail sentencing which has strong repercussions on one's life–in healthcare–such as ranking to access an organ–in education–such as enrollment into a given university–in workplace–such as access to a promotion–and so on. High-stakes decisions are commonly regulated in some way by social expectations or laws, such as the GDPR [Eur16]. However, high-stakes decisions are only qualified as such when taken in the context of their application, and is not a qualifier of the process by which the decision is taken. Thus, issues regarding decisions require considering the decision system as a whole, which in the case of privacy and fairness issues often implies individuals, cohorts, and institutions.

The main focus of this thesis is therefore to look at both privacy and fairness issues in the context of algorithmic high-stakes decisions systems under the lens of both technical considerations and social obligations. A parallel is then drawn between this unification of concepts and the current state of fairness in computer science, with many models and definitions. These limitations are already heavily discussed in the humanity communities, as illustrated by the following quote from [BP08]:

> *The system of inequalities is characterized not only by the interactions between all its constituent elements, but also by the retroactive nature of some of these interactions. This feedback can be positive, with the effect reinforcing its own cause, which will in turn reinforce it in a cumulative process; or, on the contrary, negative, leading in this case to the establishment of a certain equilibrium by self-regulation.*

However, the computer science and particularly the machine learning community have not really researched thoroughly this topic.

Privacy

In privacy protection, the first responses were syntactic models, transforming the data to providing some sort of record indistinguishability [CT13]. For example, k-anonymity [Swe02], introduced in 2002, hides individual profiles by grouping them into clusters of at least one person. This model of privacy has been successively refined to control utility–more precisely, information loss–and to respond to attacks, leading to l-diversity [Mac+07] and then t-closeness [LLV07].

In parallel, differential privacy [Dwo06] was proposed in 2006 with the objective of proposing a more robust definition of privacy. More precisely, a mechanism respects

privacy–in the sense of differential privacy–if the presence or absence of an individual in the database provided as input to the mechanism impacts the calculation in an indiscernible way, obeying to an allocated budget, which is used to bound the privacy loss, that is, the amount of information that can leak In other words, given the result of a query, an attacker gains only limited information to infer the presence or absence of the individual.

If this definition is considered, in this work as in many others, as the reference, a great many variations have been proposed [DP20] in order to respond to different models of adversary, auxiliary knowledge or other data structures. To the issue of adopting a definition is added that of choosing a budget–and additional variables when considering variations-with attempts at academic recommendation [LC11; Hsu+14] and different practical uses in industry [Tan+17].

Fairness

Non-discrimination, or fairness, is a more recent topic in computer science than privacy, but there are similarities in the approach taken to address it. Indeed, the many ways of conceptualizing discrimination and its resolution have materialized in multiple algorithmic redefinitions of complex notions. The most common definitions are the so-called "group fairness", which mostly take as their framework the decisions made by supervised learning algorithms, seeking to compare the output of a system–i.e., its decisions–with ground truth data. Within this framework, group discrimination is defined as the difference between a certain statistical indicator, such as the rate of positive decisions or the error rate, evaluated between two groups. This notion of group is defined through so-called protected attributes–such as race, religion, gender, etc. A sensitive group is then a group, identified by one or multiple protected attributes, which is discriminated against. It is to be distinguished from the concepts of statistical majority and minority Many definitions belonging to this family exist [Meh+22], with some of them being mutually exclusive [FSV16].

Another proposition is "individual fairness" [Dwo+12], which requires that two "similar" profiles receive similar decisions, given some notion of similarity between individuals. The definition of fairness then relies on this measure of similarity, which must itself be non-discriminating. It is interesting to note that Cynthia Dwork, the first author of this work, is also behind the definition of differential privacy. We can also cite definitions based on causal models [PB22], in which the aim is to remove the impact of confounding variables in the decision-making process, which are variables influencing both the decision and the input variables.

This multiplicity of definitions and concepts is similar to some extent to what can be observed for privacy, with the difference that no definition emerges as technically or socially preferable as does differential privacy in privacy protection. In particular, it seems neither desirable nor possible to achieve such convergence, both because of the formal incompatibilities [FSV16] of "desirable" concepts, and because of the dependence between interest in a concept and the agent who formulates it [Nar18], leaving room for

bias in the choice of the definition itself.

As an example we can take the intersectional analytical framework [Cre89], as coined by the American Afrofeminist scholar Kimberlé Williams Crenshaw, which allows to consider the various forms of domination, oppression and discrimination not as compartmentalized, but as an interconnected system, with the idea that these are "greater than the sum of their parts" [PB20]. While this notion is described and studied at length in several works in the human science community, it is formalized in multiple technical articles [Mor+19; Fou+20; Kan+22; ML22] as the *simple* consideration of a multiplicity of sensible attributes, such as the intersection of gender and race. Although these proposed formulations provide a first step towards the use of a mainly abstract concept in a technical environment, the reduction to inequalities, for example formulated on the model of differential privacy and applied to subgroups of the population in [Fou+20], cannot retain the context and methodological principles of the approach. The complexity of using this framework is illustrated by [Bau+21] who compares it across several quantitative studies, or by [VK22], who proposes a first empirical definition.

Contributions

To investigate the fundamental issues raised by the technicization of social challenges, and as an attempt to bring these sociotechnical challenges back into their social context, this thesis tackles a broad set of questions in both privacy and fairness.

Privacy contextual issue The use of privacy-preserving mechanisms can be hampered by the context in which data is used. Examples at the technical level are undefined context such as with natural language, which makes background knowledge hard to quantify, or unstructured data which requires alternative privacy definitions and mechanisms. Natural language is indeed full of implicit knowledge and cultural bias which are difficult to encode in language model, rendering the background knowledge difficult to quantify. Additionally, contextual requirements such as strong utility requirements or output constraints can pose challenges greater than the common privacy trade-off. A first problem we study is therefore how formal privacy models can be used under strong societal constraints on their behavior and output properties. In particular, we look at the privacy-transparency trade-off faced when publishing legal proceedings, and ask ourselves how to navigate the trade-off between privacy and utility when publishing legal decisions.

To answer this first question, we examine in a first contribution [ABG20b; ABG20a] the conflict between privacy and transparency in the publication of legal proceedings, we illustrate the inadequacy of generic methods and the difficulty of their application under specific constraints. Court decisions are now published massively online, both for the sake of transparency and accessibility of justice, and to meet the demand for legal technologies building on the wealth of legal data available. However, the sensitive nature of legal decisions also raises important privacy issues. Current practices solve the resulting trade-off between privacy and transparency by combining access control

with text redaction. We show that current practices are insufficient for coping with massive access to legal data–restrictive access control policies is detrimental to openness and to utility while text redaction is unable to provide sound privacy protection–and advocate for a integrative approach that could benefit from the latest developments of the privacy-preserving data publishing domain. We propose to consider transparency as a human desideratum on the one hand, and privacy as a requirement for automated processing on the other. Discarding the one-size-fits-all approach allows to circumvent the tensions between privacy and transparency. To this end we propose a straw man multimodal architecture paving the way to a full-fledged privacy-preserving legal data publishing system.

Privacy in practice Formal privacy models and mechanisms often depend on parameters which control their guarantees. However, selecting these parameters is challenging and often depending on contextual, cultural and individual considerations. Although following common practice, adopting economical frameworks [Hsu+14], or using risks based approaches can be a first step to guide the selection of a model and its parameters, understanding the actual guarantees offered by models and mechanisms still depends on many factors. The second problem we study is to how to address this issue of evaluating privacy mechanisms in practice.

To this issue, In a second contribution [ABG23c], we propose a framework for organizing attack competitions of privacy-preserving data publishing mechanisms in order to better establish their behaviors and limits in practice. Recently, several challenges focusing on privacy-friendly publishing mechanisms were organized, improving the understanding of trade-offs between privacy and utility, and leading to the implementation of state-of-the-art or even new sanitization algorithms. Although some challenges integrate an attack phase, their results have generally been more mixed, such as equivalent to random guesses in one case [Jor+20]. We aim to focus participants' energy and expertise against a small set of carefully chosen state-of-the-art sanitization algorithms, giving participants sufficient time to carefully design and test their attacks. To this end, we propose a comprehensive and generic system adaptable to different attack frameworks, and we define a set of abstractions in the form of a graph of rules for transforming input data and participant submissions into their evaluation. Using this framework, we organize an international attack competition [ABG23b] on privacy-preserving tabular synthetic data generation mechanisms, collocated with the French 13$^{\text{ième}}$ Atelier sur la Protection de la Vie Privée.

Observing (un)fairness Most definitions of fairness focus on the algorithmic part of the process by evaluating the fairness of the algorithm rather than that of the decision. That is, current fairness metrics allow the producer of decisions (i.e., the maker of a model) to guarantee that the model *do not add unfairness* (with respect to some definition), not that it *is fair*. This better fits a legal requirement than an ethical/societal goal of fairness. Indeed, group fairness definitions compare the output of the algorithm (the decision) to the input (the training data) and assess whether the algorithm introduced

bias, not whether the output *is biased* is the first hand. These issues result in the absence of an "absolute fairness" notion, due both to the conceptual goal of decisions producers–who only need their *product*, (i.e., the model), to be fair–and a technical limitation of existing propositions being mutually exclusive. Furthermore, this results in "snapshot" of fairness notions, which offer a restricted consideration of the impact on the system and its individuals. For instance, this notion focuses on the decision itself and ignores the way this decision can be used in the future by other algorithms in conjunction with other information. To address this shortcoming of a lot of works in the fairness literature, we ask ourselves how to consider the long-term effect of deploying fairness-aware or fairness interventions methods.

To this end, in a third contribution, we focus on the limits of technical definitions of fairness and use a reality-based simulation as a means of observing their long-term impact on a system as a whole. The definitions of fairness used today often propose a "snapshot" conception in which the algorithm's output is considered without taking into account its future impact, i.e., the context in which the prediction can be used to make other decisions. As a first step towards considering the impact of a decision and of fairness constraints on individuals, groups and the system as a whole, we propose to use a simulation grounded in reality. To do so, we start by presenting our effort towards reproducing an educational policy model [ABG23a] which we use to evaluate the behavior of technical fairness in a context grounded in reality. This technical tool meets the need, albeit imperfectly, to observe the behavior of non-discrimination mechanisms without requiring immoral experimentation on individuals or impossible experimentation on a society. We adapt this mechanism to take into account two dependency mechanisms: the reinforcement loop for decision-making systems and the "snowball effect" for individuals, which consists in considering an accumulation of decisions from the point of view of individuals: a decision received is potentially then used to make another decision, as is the case, for example, for the move from a high school to a university. From this model, we observe the behavior of simple non-discrimination mechanisms and how their effect is measured or measured by different metrics.

Outline

We first introduce the background in Chapter 2. In Chapter 3 we present our first contribution, in which we study the tension between transparency and privacy when publishing court decisions. Chapter 4 describes and demonstrates the practicality of a framework designed to facilitate the organization of attack challenges. We present our third contribution in Chapter 5, in which we study the long-term dimension of fairness. We conclude in Chapter 6.

2. Background

A background on privacy and fairness–though these topics should obviously be on the foreground.

2.1. Privacy

2.1.1. Historically many models

One of the first popular technical definition of privacy is k-anonymity [Swe02], first conceptualized in 1986 [Dal86]. It relies on the idea of hiding single individuals into groups, requiring that the information of each person cannot be distinguished from at least $k-1$ other individuals. This means that each combination of quasi-identifiers–such as the combination of gender, birthdates and postal codes [Swe00]–must be present at least k times. Generally, to achieve k-anonymity, a space partitioning algorithm is applied to the data to compute generalized values, which are used in place of the real values for each individual. Figure 2.1 is an example of such a decomposition computed by following the Mondrian algorithm [LDR06]. The representation of values depends on the context and type of each column. When a taxonomy is available, categorical values can be generalized into a more general topic, while numerical values are typically replaced with a range or average value.

Multiple limitations of the k-anonymity model have been described. Without going into details, we describe the particular "complementary release attack" [Swe02], which will be particularly relevant when discussing differential privacy in the next section. When releasing data adhering to k-anonymity, quasi-identifiers are selected with respect to other external available information. Subsequent related releases must therefore consider all the released attributes as quasi-identifier to prevent linking. Similarly, if quasi-identifiers are wrongly defined then the guarantee falls, such as if unknown previous releases exist. The same issue arises if an attacker has access to background knowledge, which could be used as new quasi-identifiers.

The model of l-diversity [Mac+07] was proposed to refine the definition and specifically counter the background knowledge made of negative sentences (e.g., "Bob does not have cancer") attacks and other attacks benefiting of cases where sensitive attributes have little diversity. It builds upon k-anonymity and adds intra-group diversity, requiring that classes contain at least l "well-represented" sensitive attributes, in which "well-represented" can be defined in several ways such has simply existing, or having a minimal entropy.

A third refinement to the definition is t-closeness [LLV07], which considers the loss of utility incurred by k-anonymity and l-diversity, which takes into account the distribution

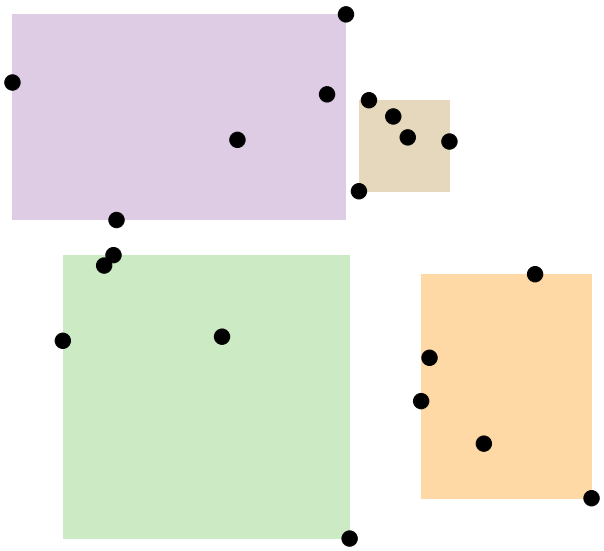


Figure 2.1.: Example of Mondrian execution with $k = 5$ using the two-dimensional plane as quasi-identifiers.

of values for the attributes. This representation also helps with an issue of l-diversity, which does not consider semantic closeness when enforcing diversity.

While we will not be using these privacy definitions in the rest of this thesis, it is worth presenting them to review and discuss the analogy we make between the development of the privacy research and the development the fairness one.

As a key takeaway, it is interesting to take into account the incremental aspect of the development of privacy notions, and it was motivated by attacks against previous contributions.

2.1.2. Differential privacy

The approaches presented above define privacy as a property of the (sanitized) dataset, which can lead to information leakage. On the other side, differential privacy is defined as a property of the publication mechanism. It was inspired [DP20] by the privacy desideratum of Dalenius that "anything about an individual that can be learned from the dataset can also be learned without access to the dataset" [Dal77].

Feminist perspective Since we consider the social dimension of privacy, it is interesting to note that women have been especially present in the field, relative to other domains in computer science. Although not backed up by studies to our knowledge, this relative over-representation might be an indicator of the specificity of social issues. This was noted by Edwige Cyffers and discussed in [Ben23], giving the prominent examples of Helen Nissenbaum, Cynthia Dwork, and Latanya Sweeney. [The+21] also provides a feminist perspective of the field of privacy. The same observation can be made for fairness, with discriminated communities having a strong interest and being

concerned by the issue, as highlighted by initiatives such as Queer in AI [DM23].

Today's gold standard Differential privacy is used by many organizations. It was adopted by the U.S. Census Bureau for the 2020 Census [Abo18], by Google as a way to report usage statistics in its Chrome web browser [EPK14], by Apple to do analytics on emoji usage [Tan+17], by Microsoft for its telemetry [DKY17], or by the Google to anonymize statistics on the response to policies aimed at combating COVID-19 [Akt+20].

Differential privacy is defined as a property of a mechanism. In a nutshell, ϵ-differential privacy ensures that the presence or absence of the data of a single individual has a limited impact on the output of the computation, thus also bounding the inference that can be done by an adversary about a particular individual based on the observed output [ABG20b].

Definition 1 (ϵ-differential privacy [Dwo06])**.** A randomized mechanism M satisfies ϵ-differential privacy with $\epsilon > 0$ if

$$\Pr[M(\mathcal{D}_1 = \mathcal{O})] \leq \exp(\epsilon) \cdot \Pr[M(\mathcal{D}_2) = \mathcal{O}]$$

for any $\mathcal{O} \in \mathit{Range}(M)$ and any tabular dataset $\mathcal{D}_1$ and $\mathcal{D}_2$ that differs in at most one row, in which each row corresponds to a distinct individual.

Differential privacy exhibits a set of composability properties that helps to analyze the impact on the overall privacy guarantees of using a differentially-private scheme.

Definition 2 (Sequential and parallel composability [DR14])**.** Let M_i be a set of mechanisms such that each provides ϵ_i-differential privacy. First, the sequential composability property of differential privacy states that computing all mechanisms on the same dataset results in satisfying $(\sum_i \epsilon_i)$-differential privacy. Second, the parallel composability property states that computing each mechanism on disjoint subsets provides $(\max_i \epsilon_i)$-differential privacy.

Definition 3 (Post-processing [DR14])**.** The post-processing property of differential-privacy states that it is always safe (i.e., at no risk of breaking the privacy guarantee) to perform arbitrary computations on the output of a differentially private mechanism. Formally, if M satisfies ϵ-differential privacy, then for any mechanism F, $F \circ M$ satisfies ϵ-differential privacy.

Answering numerical queries Since differential privacy is a property of the mechanism rather than the data, its implementation depends on the actual query to be considered.

A common approach to implement differential privacy is to use the Laplacian mechanism, which adds noise drawn from a carefully parameterized Laplace distribution to the true answer to a numerical query. We introduce this mechanism on count and sum queries as a way to illustrate the impact of the budget ϵ.

Definition 4 (Laplace mechanism [DR14])**.** Given a numerical function $f\colon X \to Y \subset \mathbb{R}$, the function $F\colon X \to \mathbb{R}$ defined as follows satisfies ϵ-differential privacy:

$$F(x) = f(x) + Laplace\left(\frac{\delta}{\epsilon}\right)$$

in which δ is the sensitivity of f and $Laplace(s)$ denotes sampling from the Laplace distribution with center 0 and scale s.

The sensitivity of a function is the maximal amount by which the output changes when a profile is removed or added to the original dataset. For example, adding an individual to a dataset will change the result of a count query by one, whereas the sensitivity of a sum query will depend on the domain of the values.

Randomized response Another example of a mechanism is *randomized response* [War65; Gre+69]. It provides a simple example of using differential privacy, and–benefiting from its simplicity–also illustrates the need for privacy when considering social issues. First proposed by Stanley L. Warner in 1965–and subsequently redefined to increase its utility–it allows a *single individual* to *locally* anonymize its answer to a sensitive query. It was used to survey sensitive topics such as cheating to exams [SD87], drug usage [GG75], or abortions [AGH70], which all require respondents to be confident in the plausible deniability of the scheme. The way noise is applied locally, and therefore before the aggregation, results in a high loss of utility, which is one of the drawback when using local differential privacy [Kas+11].

Many variations Although differential privacy is often presented–including by this thesis–as *the* flagship privacy definition, many variations of the model have been proposed. For instance, [DP20] lists a hundred of such variants and extension of differential privacy, organized on seven dimensions. These variants allow data custodian to consider different risk models or contexts of application.

Privacy washing With the increase in privacy risks, privacy protection becomes a marketing factor. Similar to the concept of *greenwashing*, which is a form of advertising in which marketing is deceptively used to persuade that products, aims and policies are environmentally friendly, privacy washing is a recent phenomenon in which privacy (often along with security) awareness is made into a selling point without actual consideration of the issue. This is made easier by the technical knowledge required to evaluate an approach. Indeed, weak mechanisms can be used, or strong mechanisms misused. An example is the case of Apple discussed before, in which the "bold announcement" [Tan+17] of its deployment of differential privacy was criticized for its lack of transparency regarding the implementation and atypical parameters that lead to low privacy guarantees [Tan+17].

2.1.3. Privacy attacks

Privacy attacks of decision systems, and usually of machine learning models, are categorized into different types. For instance, [RG20] proposes a taxonomy composed of reconstruction, property inference, model extraction, and membership inference attacks.

The aim of a *reconstruction attack* [FJR15; HZL19; Zha+20] (sometimes called attribute inference or model inversion) is to recreate part of the training data of a model. Some variations create class representatives rather than an actual reconstruction of the data.

Property inference attacks [Ate+15; SR20] focus on the extraction of information that was not encoded in the dataset, that is, to infer a property on the whole dataset or on a subgroup of individuals. For instance, [RG20] gives the example of a model that "performs gender classification and can be used to infer if people in the training dataset wear glasses or not", which can interestingly be linked to fairness issues.

A *model extraction attack* [Tra+16; Kri+20] considers black-box models–that is when the adversary has no knowledge of the model parameters, architecture, or training data–and attempts to extract parameters and potentially reconstruct a substitute model that behaves similarly to the model under attack. This attack can be used as a first step towards other attacks, such as membership inference attacks [NSH19] described hereafter.

A *membership inference attack* aims at determining whether a sample (i.e., an individual in our case) was used as part of the training data. As noted in [RG20], this is a popular category of attacks, with more than 30 examples referenced in [Hu+22] It was introduced in the machine learning setting by Shokri et al. [Sho+17]. The actual privacy risk of such an attack depends on the nature of the data at hand. For example, we can consider a model trained to predict a patient's medical procedure, which could then be attacked to infer whether a particular individual was hospitalized. It is interesting to note that the definition of membership inference is closely related to the definition of differential privacy, which is why it is often used as an empirical evaluation of the lower bounds of privacy guarantees in privacy auditing papers [NSH19; JUO20].

We focus on this specific type of attack in the rest of this section, describe the initial attack as proposed by [Sho+17], and how its success can be evaluated.

Shadow models attack The attack described by [Sho+17] and illustrated on Figure 2.2 is now a common design pattern for a lot of membership inference attacks [Hu+22]. It is based on the intuition that a model will behave differently when operated on data that do not belong to the previously seen training data. This can be captured either in the model output or in the internal model representation, which the attack model then uses to distinguish members from non-members.

The actual attack model contains a collection of models, one for each output class of the model under attack, called "shadow models". Since we consider here binary decisions, only two models are required. To build the attack, we consider a number of datasets D_{shadow} that are usually assumed to come from the same distribution as the target dataset. Each dataset D_{shadow} is partitioned into sets D_{in} and D_{out}, and a model M_{shadow}–typically with the same architecture as the target model–is trained on

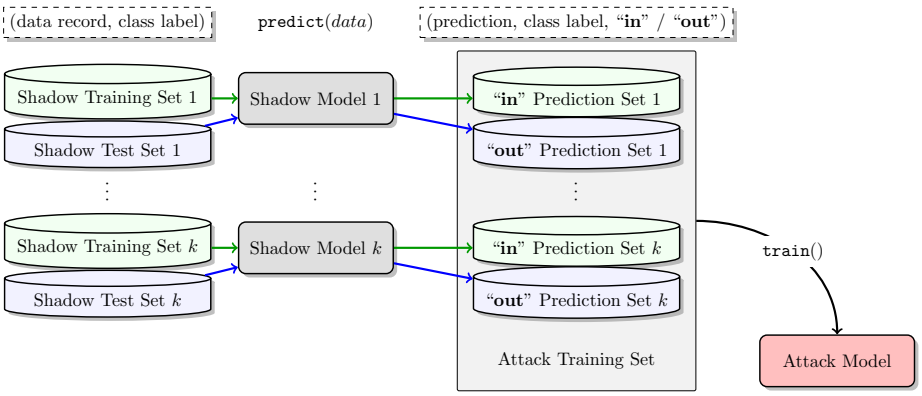


Figure 2.2.: Training of the shadow model attack, reproduced from [Sho+17]

D_{in}. The samples $(\boldsymbol{x}, y) \in D_{in}$ are used to build an "in" prediction set composed of samples $((y, \boldsymbol{y} = M_{shadow}(\boldsymbol{x})), in)$. Similarly, D_{out} is used to build a dataset of samples $((y, \boldsymbol{y} = M_{shadow}(\boldsymbol{x})), out)$. These samples form the attack training set D_{inout}, which is used to train a metamodel A. Given a target sample $(\boldsymbol{x}, y)$, the attacker gets $\boldsymbol{y}$ by feeding $\boldsymbol{x}$ to the target model, before predicting whether the data point is in the training set by evaluating $A(y, \boldsymbol{y})$.

Membership advantage The evaluation of membership inference attack is commonly performed by computing the *membership advantage* which estimates the knowledge gained by the attacker from the release, as compared to a guess done only from prior knowledge [Ye+22].

Definition 5 (Membership experiment $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$ [Yeo+18])**.** Let $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, and $\mathcal{D}$ be a distribution over data points (x, y). The membership experiment proceeds as follows:

1. Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
2. Choose $b \leftarrow \{0, 1\}$ uniformly at random.
3. Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$
4. $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$ is 1 if $\mathcal{A}(z, A_S, n, \mathcal{D}) = b$ and 0 otherwise. $\mathcal{A}$ must output either 0 or 1.

Definition 6 (Membership advantage [Yeo+18])**.** The membership advantage of $\mathcal{A}$ is defined as

$$\mathsf{Adv}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D}) = 2 \ \Pr[\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D}) = 1] - 1, \tag{2.1}$$

in which the probabilities are taken over the coin flips of $\mathcal{A}$, the random choices of S and b, and the random data point $z \sim S$ or $z \sim \mathcal{D}$.

Equivalently, the right-hand side can be expressed as the difference between $\mathcal{A}$'s true and false positive rates

$$\mathsf{Adv}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D}) = \Pr[\mathcal{A} = 0 | b = 0] - \Pr[\mathcal{A} = 0 | b = 1]. \tag{2.2}$$

2.2. Fairness

The first step required to define fairness in a technical setting is to have a concept of discrimination. This can originate from legal definitions, cultural norms, ethical philosophy, and so on. Many such propositions exist, and are outside the expertise of this thesis. A second issue is to have a notion of group on which to evaluate the discrimination. These groups are then refereed as protected and defined by one or multiple protected attributes. Typical protected attributes are race, gender, religion, etc. and can again vary across societies. In the next section, we briefly present the main families of fairness definitions encountered in the technical literature.

2.2.1. Many concepts, many frameworks, many models

A first naive approach to defining non-discrimination is to require that the decision is made while ignoring all protected attributes. However, this idea of "fairness through unawareness" is ineffective due to the existence of redundant encodings [PRT08], similar to the issue with quasi-identifiers in privacy.

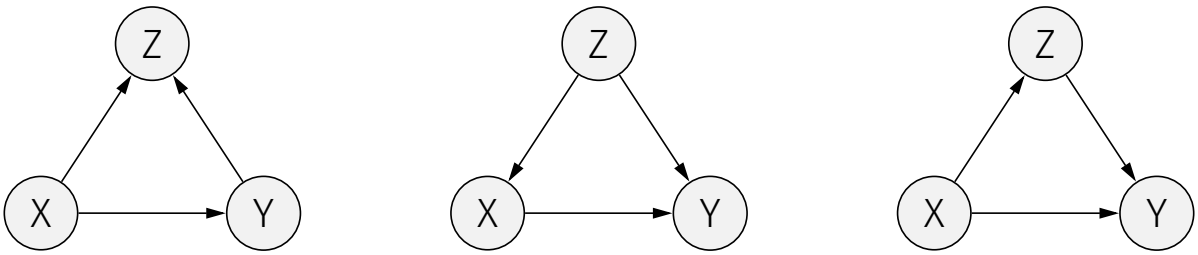
Many fairness definitions have been proposed [Meh+22; GB20], with two main families: individual and group-based.

Individual fairness [Dwo+12] was proposed by Cynthia Dwork. It requires that similar individuals receive similar outcomes, for some similarity metric which itself needs to be fair. The key idea as defined by [GK21] is to introduce a Lipschitz condition (i.e., a uniform continuity) on the decisions of a classifier, such that for any two individuals x, y that are at distance $d(x, y)$, the corresponding distributions over decisions $M(x)$ and $M(y)$ are also statistically close within a distance of some multiple of $d(x, y)$.

The group fairness family contains many metrics. At an abstract level, it requires some statistical indicator to be equal (or similar) across groups. These indicators are commonly expressed through a combination of variables from the confusion matrix when considering supervised learning. Common examples include equalized odds [HPS16], which requires subjects in the protected and unprotected groups to have equal true positive rate and equal false positive rate, and equal opportunity, in which subjects in the protected and unprotected groups must have equal false negative rate.

Some definitions of fairness are formulated from privacy concepts, requiring that an attacker cannot infer the protected attribute from the output of a system [Fel+15].

Finally, causality-based fairness notions are a more recent proposition [PB22]. They differ from the statistical ones such as group fairness in that they consider additional knowledge about the structure of the world, in the form of a causal model [CW22]. Simple causal models can be seen in Figure 2.3. Different formulations have been proposed,



(a) Collider (b) Fork (c) Chain

Figure 2.3.: Examples of connections in a causal model, with a collider, fork, and chain. In (a), Z is a collider since it is influenced by both X and Y. In (b), Z is a confounder since it impacts both the input variable X and the output variable Y. In (c), T is a mediator since it lies between the input variable X and the output variable Y in one possible path.

corresponding to different formal models of causality, as seen in [CW22]. To give a brief overview of the family, we can consider that it views discrimination as the causal effect of the sensitive attribute (Z on Figure 2.3) on the output attribute (Y on Figure 2.3). Then, multiple questions correspond to different causal fairness notions, such as total causal fairness [Pea19] which measures the impact on the output of changing the value of the protected attribute, or counterfactual fairness [Kus+17] which attempts to evaluate how the output variable would have behaved in an "alternate" scenario in which the protected variable has no impact.

It has to be noted that many papers do not fall in those categories and try to come up with their own definitions, such as [Fou+20; GB20], or mixing concepts from individual and group fairness.

Intervention Although we focus in this work on how fairness can be evaluated, we briefly review how these notions of fairness are achieved in machine learning. Fairness-aware machine learning is typically available at three levels [Meh+22; Bel+19]:

Pre-processing algorithms are used before the actual model learning by modifying the training data [KC11; Zem+13; Fel+15];

In-processing algorithms encompass many techniques where the learning step is modified to account for fairness [ZLM18; Kea+19];

Post-processing algorithms, finally, adjust the probability of specific decisions being made [Ple+17].

Long-term considerations Recent work has focused, as we do, on evaluating the long-term impact of fairness interventions. The most prominent of these attempts is ML-fairness-gym [DAm+20], which was developed as a framework to understand long-term fairness via simulation. It implements scenarios proposed in previous works [Liu+19; HIV19], and leverages OpenAI Gym [Bro+16] framework. As noted by authors of the

paper, the simulations available are "extremely simple" in both the behavior of agents and the dynamics between components. On the other hand, the focus on machine learning makes the system difficult to interpret when studying on the behavior of fairness notions rather than on particular machine learning models.

2.2.2. Practical limits

Although this thesis tackles both privacy and fairness, we did not look the intersection of the two. This occurs in several ways, which we briefly discuss hereafter even if they are out of our scope. Firstly, there is often a technical similarity between privacy and fairness. This is heavily apparent in the case of individual fairness, sharing authors and definitions. In fact, many researchers from the fairness community originate from the privacy one [Ben23]. The question of the compatibility between fairness and privacy, be it as a trade-off or as equivalent objectives, is also the subject of many studies [Cum+19]. In the vein of a trade-off, some studies have raised concerns regarding the unfair impact of using differential privacy, especially in the case of the CENSUS [Fio+22; Puj+20].

Snapshot discrimination We use the term "snapshot fairness" to refer to notions of fairness that do not consider the long-term impact, but rather are limited to enforcing a set of fair decisions at some given time. This is particularly well illustrated when considering most group fairness metrics, which evaluate a given set of decisions against a previous set of labeled decisions. The need to consider long-term dynamics of data and mechanisms with respect to their fairness, and the difficulty of identifying guidance towards these approaches was highlighted by Chouldechova et al. [CR20].

Fairwashing In line with our previous remark considering privacy washing, the same issue obviously appear with fairness, as it would with any social desiderata. In [Aïv+21], Aïvodji et al. show that given an unfair black-box model, it is possible to derive an explanation–in the form of an interpretable model–which exhibits both high-fidelity with respect to the model under question and a fairer behavior. This issue highlights the risk of trusting fairness advertising, and the difficulty of further assessments.

Fair individual fairness When looking at individual fairness, a particular difficulty–among others [Fle21; GK21]–is the reliance on a similarity measure between individuals and between treatments. The fairness rests on the fairness of this measure, which poses an obvious issue of applicability.

Inaccessible causal model Turning to the causal model, as we discussed above, it requires some knowledge of the model, which can also be a source of bias, and is difficult to obtain when considering complex sociotechnical systems. [FES22] argues that causal models used in fairness settings "entail a particularly strong causal assumption, normally only seen in a randomized controlled trial". This strong requirement is unlikely to hold and makes the applicability in real cases particularly difficult.

Formal incompatibility As a final technical issue, we go to the most obvious concern with technical fairness definitions: their incompatibility. Indeed, [FSV16] shows that, apart from unrealistic scenarios–such as rejecting everyone, most fairness metrics are incompatible between each others. This can be illustrated in the case of group fairness metrics by considering that most of these metrics are expressed from a limited set of variables available in the confusion matrix, with only unrealistic scenarios being able to satisfy more than 3 metrics. [Cho17] gives the example of predictive parity (the likelihood of positive outcome is the same across groups) and error rate balance (the false positive and false negative error rates are equal across groups), which can only be both satisfied if the prevalence is the same between groups.

At a higher level, and as discussed in [FSV16; BHN19], this also corresponds to the deeper issue of different fairness concepts corresponding to different worldviews and assumptions. This is also reflected in [Hof19], which advocates that technical fairness fails to address the "very hierarchical logic that produces advantaged and disadvantaged subjects in the first place". In other words, these approaches consider an effect–the unfair model–rather than the cause–an unfair society.

Privacy impact As a final point and to link the two topics of privacy and fairness, recent work has looked at two similarities and tensions between the two.

[Fio+22] surveys the work at the intersection of differential privacy, and fairness, showing how the two can both align or contrast, with differential privacy being able to sometimes mitigate fairness issues, or to exacerbate disparities among groups of individuals. In [Puj+20] the issue is illustrated on the U.S. Census data where the use of differential privacy disproportionately impact some groups over others for, as an example, the assignment of voting rights benefits. [CS21] uses membership inference attacks as a framework to quantify such disparate impact.

3. The privacy-transparency trade-off of legal decisions publication

This chapter is adapted from [ABG20b; ABG20a], and was presented as an invited talk at the webinar on the use of AI in the justice field organized by the European Commission in 2021. First in Section 3.1 we introduce the context under which legal decisions are published, and the associated utility, transparency and privacy constraints. In Section 3.2 we state the problem by describing precisely the content of legal data and by explaining the open legal data desiderata. Afterwards, we present the privacy limitations of the current approach, redaction, in Section 3.3.2, before describing in Section 3.4 our proposal of architecture for the publication of legal data ensuring both privacy and utility.

3.1. Publication of court decisions

The opening of legal decisions to the public is one of the cornerstones of many modern democracies: it allows auditing and make accountable the legal system by ensuring that justice is rendered with respect to the laws in place. As stated in [BB43], it can even be considered that "publicity is the very soul of justice". Additionally, in countries following the common law, the access to legal decisions is a necessity as the law in place emerged from the previous decisions of justice courts.

Thus, it is not surprising that the transparency of justice is enshrined in many countries as a fundamental principle, such as the *right to a public hearing* provided by the Article 6 of the European Convention on Human Rights, the Section 135(1) of the Courts of Justice Act (Ontario) stating the general principle that "all court hearings shall be open to the public" or in Vancouver Sun (Re) "The open court principle has long been recognized as a cornerstone of the common law". The open data movement push for free access to law with for example the Declaration on Free Access to Law [Wor02]. Multiple open government initiatives also consider the need for an open justice, such as the "Loi pour une République numérique" in France, the Open Government Partnership, the Open Data Charter, the Canada's Action Plan on Open Government. This trend is studied in a report of the OCDE [Org11], in [McD10] for the USA or [McC11] for the UK.

Combined with recent advances in machine learning and natural language processing, the (massive) opening of legal data allows for new practices and applications (called legal technologies). Nonetheless, not all legal decisions should directly be published as such due to the privacy risks that might be incurred by victims, witnesses, members of the

jury and judges. Some privacy risks have been considered and mitigated by legal systems for a long time. For instance, the identities of the individuals involved in sensitive cases, such as cases with minors, are usually *anonymized* by default because they belong to a vulnerable subgroup of the population. In situations in which the risks of reprisal are high (e.g., terrorism or organized crimes cases), judges, lawyers and witnesses might also ask for their identities to be hidden [Jac17; Fle17]. Finally, the identities of the members of a jury are also usually protected to guarantee that they will not be coerced but also to ensure that the strategy deployed by the lawyers is not tailored based on their background. Legal scholars are aware of the need for privacy when opening sensitive legal reports [Con+11; Jac02; BB16].

In the past, these privacy risks were limited due to the efforts that were required to access the decisions themselves. For instance, some countries require going directly to the court itself to be able to access the legal decisions. Even when the information is available online, the access to legal decisions is usually on a one-to-one basis through a public but restricted API rather than enabling a direct download of the whole legal corpus. Typical restriction mechanisms include CAPTCHAs (SOQUIJ), quotas (CanLII), registration requirement as well as policy agreement and limitation of access to research scholars (Caselaw). Furthermore, the fact that a legal decision is public does not mean that it can, legally, be copied and integrated in other systems or services without any restrictions.

A first approach to limit the privacy risks consists in *redacting* the legal decisions before publishing them. Redaction mostly follows predefined rules that list the information that must be removed or generalized and define how (e.g., by replacing the first and last names by initials, by a pseudonym) [Opi+17]. Redaction is in general semi-manual (and sometimes fully manual) because automatic redaction is error-prone [Mar+13]. This makes it extremely costly, not scalable, and does not completely remove the risks of errors [Opi+17]. For example, 3.9 million decisions are pronounced in France every year but only 180000 are recorded in government databases and less than 15000 are made accessible to the public [Fou+19]. Moreover, even a perfect redaction would still offer weak privacy guarantees. A redacted text still contains a non-negligible amount of information, possibly identifying or sensitive, that may be extracted, e.g., from the background of the case or even from the natural language semantics.

Another approach is access control, such as non-publication (e.g., a case involving terrorism was held in secret in Britain [Cal14]), rate limit, or registration requirements. However, access control mechanisms are binary and do not protect the privacy of the texts for which the access is authorized. Furthermore, restricting massive accesses for blocking also restricts the development of legal technologies that require a massive access to legal data.

3.2. Problem statement

3.2.1. Legal data

Legal reports are defined as written documents produced by a court about a particular judgment, which is itself a written decision of a court regarding a particular case (oral judgments are transcribed). Although the content of a case report varies with respect to the court and the country, it can consist of elements such as [Uni09]:

1. the case name and case citation (identifier);
2. the date of judgment and the hearing dates;
3. the court and judges involved in the decision;
4. the appearances (parties and their representatives);
5. the statement of facts: identify–sometimes in great length–the relationship and status of the parties, the legally relevant facts (i.e., what happened), and the procedurally significant facts (e.g., cause of action, relief request, raised defenses);
6. the procedural history: describes–if applicable–the disposition of the case in the lower court(s), the damages awarded, the reason for appeal, etc.;
7. the issues: point of law in dispute;
8. the law of the case: elements of law that the court applies;
9. the concurring and/or dissenting opinions (of judges);
10. the orders: the decision itself.

We can broadly distinguish three different categories of judicial data depending: metadata, facts and reasoning. Metadata (elements 1, 2, 3 and 4) correspond to identifiers of the case and basic information (e.g., date, parties and judge) and is written mostly in a structured way. Facts (elements 5, 6 and 7) are information pertaining to the parties, disclosing their personal "story". Reasoning (elements 8, 9 and 10) is the logic of the case, which is not specific to the parties.

3.2.2. Desiderata for the opening of legal data

In this section we present the needs that arise with the publication of legal decisions. We start by presenting utility requirements for individuals in Section 3.2.2, and for algorithms in Section 3.2.2. Finally we discuss the need for privacy in Section 3.2.2.

Need for readability and accessibility

The access to legal decisions is required both for ethical (transparency) and practical reasons such as case law, which is the use of past legal decisions to support the decision for future cases. Thus, the judiciary system is built on the assumption that legal decisions are made public and accessible by default (*open-court principle*), so that (1) citizens are able to inspect decisions as a way to audit the legal system and (2) past decisions can be used to interpret laws, and as such must be known from legal practitioners and citizens. It follows that decisions must be made available in a form readable by humans (i.e., natural language). Natural language format can be opposed to machine-readable formats such as word-vectors representation or logical propositions, which we will discuss later. The need for openness, the current practice in terms of open court, and the associated risks are detailed in [Con+11; Mar08]. They conclude that, although there are powerful voices in favor of open court, radical changes in access and dissemination require new privacy constraints, and a public debate on the effect of sharing and using information in records.

Accessibility is also an important issue. In the past, the access to decisions required attending public hearings or reading books called "reporters". Later, decisions have started to be shared on digital medium such as compact discs or DVDs for example before being accessible online more recently. For instance, in the USA, CourtListener [Fre23] shares 3.6M decisions and the Caselaw access project [Pre23] 6.7M unique cases; the Canadian Legal Information Institute [Fed23] (CanLII) publishes 2.5M Canadian decisions. The aim of these services is to facilitate access to legal records to individuals–law professionals (judges, lawmakers and lawyers), journalists, or citizens. The online publication also enables the large-scale access and processing of records, in particular due to the standardized format.

Need for massive accesses (legal technologies)

The term *legal technologies* broadly encompasses all the technologies used in the context of justice. The website CodeX Techindex [Cod23], a project by the Stanford Center for Legal Informatics, references more than a thousand companies, and defines nine different categories: (1) Marketplace, (2) Document Automation, (3) Practice Management, (4) Legal Research, (5) Legal Education, (6) Online Dispute Resolution, (7) E-Discovery, (8) Analytics and (9) Compliance.

A subset of these categories–2, 4, 7, 8 and 9–requires some form of "understanding" of legal documents, usually performed through natural language processing (NLP) and machine learning (ML) approaches [PPC16; Dal19]. We focus here on these categories as they are based on the analysis of a large number of legal data. One of the main challenges we have faced is that usually companies provide very few technical details about their actual processing and usage of legal documents.

The automatic processing and analysis of legal records have multiple applications, such as computing similarity between cases [TKA17; MS18; Man+17], predicting legal outcomes [Ale+16; KBB17] (e.g., by weighing the strength of the defender arguments

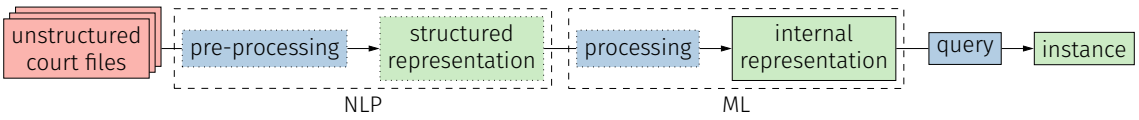


Figure 3.1.: High-level pipeline of court files processing for Legal Techs

and the legal position of a client in a hypothetical or actual lawsuit), identifying influential cases [Mar+19; Moz+05] or important part of laws [MSC19], estimating the risk of recidivism [Tan+18], summarizing legal documents [MFL10], extracting entities (e.g., parties, lawyers, law firms, judges, motions, orders, motion type, filer, order type, decision type and judge names) from legal documents [QG10; Cus+19], topic modelling [NM08; AB09], concept mapping [BA97] or inferring patterns [Kor65; AW13].

Focus on text-based legal techs Most of the technologies introduced in the previous section rely on the processing of large database of legal data. However, the unstructured nature of legal data is one of the main challenges of the application of artificial intelligence in law [ANY17]. Consequently, the analysis of a legal text corpus first requires to apply some pre-processing to add structure to the text. Figure 3.1 represents an abstract processing pipeline for court files, extracted mostly from academic papers[1], and inferred from the current practice of text analysis and descriptions of associated technologies. In the following, we assume that any application involving the use of machine learning (as highlighted by most legal tech companies) is applied to court records. The first NLP step transforms the unstructured data (i.e., natural language) into some structured representation (see below) by pre-processing it.

Afterwards, the second ML step corresponds to the actual application, which is the training (i.e., processing) of the ML algorithm, whose output is represented by the "internal representation" block. The term instance represents the output of the model given some query (e.g., applicable laws given a set of keywords representing infractions).

The pre-processing can be diverse and depends on the task (e.g., extracting a citation graph between cases). However, most NLP-based applications usually rely on a text model. Many models are statistical-based ones, such as document-word-frequency matrix, in which the corpus is decomposed into a matrix in which each cell contains the number of times a particular word appears in a document. This model has multiple variations such as bag-of-words (BoW) [Joa98], term frequency-inverse document frequency, or n-grams [ZZL15]. For example, a combination of those techniques are used in [Ale+16] to predict decisions from the European Court of Human Rights, and by [KRG19] to identify law articles given a query or to answer to questions given a law article. More recent approaches follow a neural network architecture in which a model is trained on the corpus with the objective to predict a word given a context, which is called word embeddings [Mik+13]. Multiple variations of this structure exist [Jou+17; Kim14; LQH16; Lai+15; Zhe+18]. This approach has been used for example in [Mar+19] to

[1]The majority of the legal technologies market consists in commercial applications. They do not give information about their inner working and underlying techniques.

rank and explain influential aspects of law, or by [MSC19] to predict the most relevant sources of law for any given piece of text using "neural networks and deep learning algorithms".

Need for privacy

The massive opening of legal decisions for transparency and technological reasons must not hinder the fundamental rights such as the right to privacy as emphasized by current open justice laws. In particular in this setting, the privacy of a least three main actors must be protected: namely the individuals directly involved in decisions (i.e., the parties), the individuals cited by decisions (e.g., experts or witnesses), and the individuals administering the laws (i.e., magistrates).

However, the problem of publishing legal decisions in a privacy-preserving manner is a difficult one. For instance, authorship attacks [AC08] may lead to the re-identification of magistrates behind written decisions, or the presence of *quasi-identifiers*[2] within the text decisions may lead to the re-identification of the individuals involved in or cited. Famous real-life examples, such as the governor Weld's [Swe02] or Thelma Arnold's re-identification [Arr06], both based on the exploitation of quasi-identifiers, are early demonstrations of the failure of naive privacy-preserving data publishing schemes. Thus despite the fact that legal decisions are written as unstructured text, structured information can be extracted from them, including the formal argument, the decision itself (e.g., "guilty" or "innocent"), as well as arbitrary information about the individuals involved (e.g., gender, age and social relationships).

Pseudonymization schemes simply consist in removing directly identifying data (e.g., social security number, first name and last name, address) and keeping unchanged the rest of the information (quasi-identifiers included). These schemes provide a very weak protection level, as acknowledged in privacy legislations (e.g., GDPR), which has led to the development of new approaches for sanitizing personal data in the last two decades (see for instance the surveys [Che+09; ML21]). In this paper, we focus on privacy-preserving data publishing schemes providing formal privacy guarantees that hold against several publications (as required by any real-life privacy-preserving data publishing system). These schemes are based on (1) *a formal model* stating the privacy guarantees the scheme as well as *a privacy parameter* for tuning the "privacy level" that must be achieved, and (2) *a sanitization algorithm* designed to achieve the chosen model.

3.3. Analysis of current practices

In the following section, we review the current practice for legal data anonymization and privacy regulations. We also make a connection with health data anonymization techniques on which most papers rely. To be concrete, we illustrate the privacy risks through examples of re-identification attacks. Finally, we argue that rule-based anonymization

[2] A quasi-identifier is a combination of attributes that are usually unique in the population, thus indirectly identifying an individual. A typical example is the triple `(age, zip code, gender)`.

is not sufficient to provide a strong privacy protection and discuss the (formal) issues surrounding text anonymization.

3.3.1. Redaction *in the wild*

Redaction of legal data

The redaction process consists in removing or generalizing a set of predefined terms defined by law through a semi-manual process [Opi+17]. Furthermore, access to legal documents or even public hearings can be restricted in well-defined cases. The common practice is to replace sensitive terms, as defined below, by initials, random letters, blanks or generalized terms (e.g., "Montréal" becomes "Québec"). The specific set of rules regarding protected terms and the associated replacement practice can differ between countries and courthouses [Opi+17].

According to [PLP04], the following information is to be systematically removed for any person (subject to a restriction on publication), as well as for each of his or her relatives (parents, children, teachers, neighbors, employers, colleagues, school, …):

1. names,
2. date and place of birth
3. contact details (number, street, municipality, postal code, telephone, fax, email, web page, IP address),
4. unique personal identifiers (social security number, health insurance number, medical file, passport, bank account, credit card, …),
5. personal possessions identifiers (license or serial number, cadastral designation, company name, …)

In some context, the following data is also removed if it can be used to identify one of the individuals aforementioned:

7. small communities or geographic locations,
8. the accused and co-accused if their identity is not already protected by law,
9. the intervenors (court experts, social workers, police officers, doctors, …)
10. unusual information (number of children if abnormally high, income if particularly high, exceptional occupation or function).

[Con+11] present numerous examples of legislation framing the publication of specific terms and putting restriction to the *open-court principle*. For instance, it is common by default to hide the identity of victims of sexual offenses or children in youth courts. The identity of jurors and witnesses is also kept secret to avoid coercion or parties tailoring their strategy. In addition, in the USA, the fear for national security or the possible prejudice to another trial can lead to a complete ban on reporting being issued.

E.B. Petitioner v. V.I. Respondent

Judgment for Dissolution of Marriage

Quote 1: Droit de la famille–15334, 2015 QCCS 762 (CanLII), [Can15]

Paper versus digital

The main difference between paper and digital access is the "practical obscurity" of paper records on the one hand, and the easy accessibility of digital records, on the other. The awkwardness of accessing paper records stored in a public courthouse puts inherent limitations on the ability of individuals or groups to access those records. In contrast, digital records are easy to analyze, can be searched in "bulk" by combining various key factors (e.g., divorce and children) and can potentially be accessed from any computer. Thus, traditional distribution provides "practical obscurity" [Jud03], in that it is inconvenient (i.e., time-consuming) to attend the courthouse or read case reports.

Anonymization of health data

The Health Insurance Portability and Accountability Act (HIPAA) in the USA defines the security and privacy requirements of health information for both health professionals and technologies involved in health data. The search for complying with HIPPA has led to an important body of work on the redaction of health records. In particular, automated redaction or generalization of the sensitive terms defined in HIPPA generally involves domain specific named-entity recognition and generalization of terms through medical ontologies. As a concrete anonymization tool, Scrub [Swe96] uses template matching to detect sensitive terms, which are replaced with synthetic data of similar type (e.g., a name with a name, a disease with a similar disease). t-PAT [Jia+09] replaces sensitive words or phrases–recognized by an ontology–with more general terms using an early privacy-preserving data publishing model, called k-anonymity [Swe02], to preserve the privacy of patients.

3.3.2. Limitations of current approaches

Our objective in this section is to provide examples of potential attacks in order to illustrate the technical difficulties of raw text anonymization. Quotes 1, 2, 3, 4, and 5 are translated excerpts from French and Canadian opinions.

A common redaction practice is to replace names by initials as shown in Figure 1. The uniqueness of initials [Fin93] is increased by combining multiple parties, especially if the relationship between the parties is known (e.g., in a divorce case).

A combination of attributes, which can be extracted using dedicated named-entity recognition, is presented in Quote 2: names of parties, parties are divorced, date of

Katopodis v. Katopodis

SUPREME COURT OF ONTARIO

The parties were married on August 25, 1968; a daughter was born on November 30, 1972; the parties separated in April, 1977. The wife first went to see Dr. James

Quote 2: Katopodis v. Katopodis, 1979 CanLII 1887 (ON SC), [Can79]

the association Real Madrid Club de Futbol and several players of this team, Zinedine Z., David B., Raul Gonzalès B. aka Raul, Ronaldo Luiz Nazario de L., aka Ronaldo, and Luis Filipe Madeira C., aka Luis Figo

Quote 3: CA Paris, 11^{e} ch., sect. B, 14 February 2008, *Unibet Ltd c/ Real Madrid et autres*, RG n° 06/11504, GP [Leg09]

marriage, parties have a daughter, birthdate of the daughter, date of divorce, reside in Ontario near a Dr. James. This combination could be used as a quasi-identifier by a re-identification attack.

Quote 3 is anonymized according to the CNIL recommendations of 2006, which requires the last name of individuals to be replaced by its initial. However, widely available background knowledge on the "Real Madrid Club de Futbol" combined with the (real-life) pseudonyms of the "players" trivially leaks their identity.

The de-anonymization of Quote 4 relies on the text semantics instead of background knowledge. It requires the adversary (1) to identify the link (X) between "M. […] Abdel X" and "the use of the name 'X' to designate a drink", and (2) to infer that the drink is called "sango", thus leading to the conclusion that X = "sango". While this attack may not be easy to automatize due to the hardness of detecting the semantics inference, it is, however, trivial to perform for a human (e.g., by crowdsourcing it).

Similar to Quote 2, Quote 5 could be attacked through a combination of attributes and relationship (e.g., extracted with Snorkel [Rat+20]). This opinion from the Youth court involves children and, as such, follows the strictest anonymization rules of the SOQUIJ: only the year's birthdate of children is given and names are replaced by random letters. However, an adversary can extract an extensive relationship graph (see Figure 3.2), which could be matched over a relationship database (e.g., Facebook). In this case, a quasi-identifier could be the relationship graph (or parts of it).

A study [AC11] has shown that generalization-based sanitization is vulnerable to correlation attacks in which the generalized terms can be correlated with other terms, seemingly non-identifying, in order to jeopardize the effects of the generalization and

the American company Coca Cola Company markets drinks under the French trade mark "Coca Cola light sango", of which it is the proprietor; that M. […] Abdel X, relying on the infringement of his artist's name and surname, has brought an action for damages against the Coca Cola Company […] On the ground that Abdel X maintains that, as an author and screenwriter, he is entitled to oppose the use of the name "X" to designate a drink marketed by the companies of the Coca Cola group.

Quote 4: Civ. 1re, 10 April 2013, n° 12-14.525, *Sango c/ Coca-Cola*, D. 2013. 992; CCE July 2013, n° 73 [Lég13]

X, born […] 2017; Y, born […] 2018 the children and C; D the parents

Applications are submitted for X, aged 1 year, and Y, aged 2 months. The Director of Youth Protection (DYP) would like X to be entrusted to her aunt, Ms. E, until June 25, 2019. As for Y, that he be entrusted to a foster family for the next nine months. The father has two other children, Z and A, from his previous union with Mrs. F. The mother has another child, B, from her union with Mr. G.

Quote 5: Protection de la jeunesse–186470, 2018 QCCQ 6920 [SOQ18]

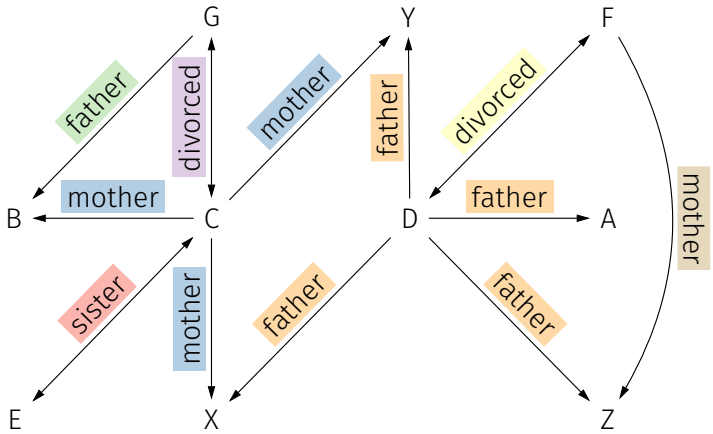


Figure 3.2.: Relationship graph manually extracted from Quote 5

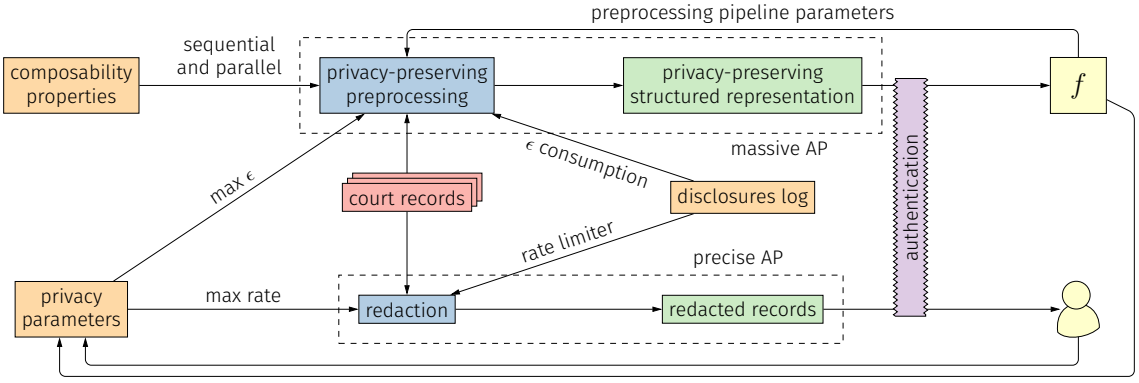


Figure 3.3.: Multimodal publication architecture

consequently disclose sensitive terms.

Besides the content of legal documents, stylometry [Nea+18] can be used to identify authors (i.e., magistrates) by their writing style. Mitigation for this kind of attack exist [FDM19; WK18] but their output is only machine readable. Similarly, it is possible to exploit decision patterns to re-identify judges, as done for the Supreme Court of the United States [KBB17].

3.3.3. Reasons for the failure of rule-based redaction

The review of current practices for tackling the privacy of legal documents in Section 3.3.1 has highlighted the widespread use of rule-based redaction, in which a set of patterns is defined as being sensitive and is either removed or replaced. However, as shown in Section 3.3.2 (1) privacy can be violated even in "simple" instances and (2) identifying information remains in most cases. In other words, rule-based redaction does not provide any sound privacy guarantee. We observe that it suffers from the following main difficulties.

1. *Missing rule difficulty.* Many combinations of quasi-identifiers can lead to re-identification and the richness of the output space offered by natural language (i.e., what can be expressed) can hardly be constrained to a set of rules. Furthermore, identifying the sensitive terms is challenging and domain-specific. This issue is the subject of multiple studies in the context of health data [Swe96; Ata+00; Dou+05].

2. *Missing match difficulty.* The current state of the art about relationship extraction and named-entity recognition makes it hard to ensure that all terms that should be redacted will be detected, in particular because of the many possible ways to express the same idea (e.g., *circumlocution*).

Although these observations make the rule-based redaction difficult, it is important to note that attacks, e.g., re-identification, remain simpler than protection. Indeed, an adversary has to find a single attack vector (i.e., a missing rule or a missing pattern) whereas the redaction process needs to consider all the possibilities.

3.4. Towards a multimodal publication scheme

In Section 3.2, we have shown that the publication of legal documents serves two distinct and complementary purposes: (1) the traditional objective of transparency and case law, and (2) the modern objective of legal technologies of providing services to citizens and legal professionals. These two purposes obey to different utility and privacy requirements. More precisely, the traditional use case requires human-readable documents while legal techs need a machine-readable format for automated processing. Moreover, transparency and case law involve the access to opinions on an individual basis (i.e., one-at-a-time), similarly to attending a hearing in person. In contrast, legal technologies rely on the access to massive legal databases. This difference in cardinality (i.e., one versus many) entails different privacy risks. In particular, the massive processing of legal data requires the use of a formal privacy framework with composability properties (see Section 3.2.2). All this suggests the inadequacy of any *one-size-fits-all* approach.

3.4.1. Access modes

As a consequence, we propose that the organization in charge of the publication of legal decisions should consider two modes of publication: the *precise access mode* and the *massive access mode*.

Precise access mode

To fulfill the "traditional" use case, the precise access mode provides full access to legal decisions that are only redacted using the current practices. This access mode is designed for the transparency and case law usages, and is to be used typically by individuals (e.g., law professionals, journalists and citizens). Similar to the "traditional" paper-based publication scheme, in the precise access mode [HS13], a user has access to text documents, either in full or only extracts (partial access is useful for crowdsourcing tagging in order to build a dataset). While the current practice of redacting identifiers could be combined with more automated approaches such as [SBV13; Has+19]. The aim of this mode is to provide strong utility first. It is similar to the websites currently publishing legal documents (e.g., Legifrance or CanLII), as it allows browsing, searching and reading documents.

To prevent malicious users from diverting the precise access mode for performing massive accesses, users must be authenticated, and their access must be restricted (e.g., rate limitation or proof of work [DN92]). The main objective of the restricted access is to make it difficult to rebuild the full (massive) database. In addition, this mode provides privacy through "practical obscurity" similarly to the paper-based system.

Massive access mode

The massive access mode gives access only to pre-processed data resulting from privacy-preserving versions of the standard NLP pipelines available on the server, i.e., aggregated and structured data extracted from or computed over large numbers of decisions,

as required for the "modern" use case. It should be compatible with most legal tech applications that traditionally use a database of legal documents (see Section 3.2.2). Note that the perturbations due to privacy-preserving data publishing schemes have usually less impact (in terms of information loss) when applied late in the pipeline (see Figure 3.3), at the cost of a loss of generality of the output.

Users need to be able to tune the pre-processing applied. For the sake of simplicity, we assume that the user (i.e., legaltech developer) provides the parameters for a given NLP pipeline (see Fig. 3.3). These parameters can be for instance the maximum number of features or n-grams range to consider in the case of a BoW model or the window for word embeddings. But more complex implementations can be designed, e.g., allowing experiments by the users, fine-tuning for each dataset/task, as well as customization (e.g., for cleaning the data). This can be done (1) by generating structured synthetic *testing* data (e.g., a set of features extracted from legal data) in a privacy-preserving manner (e.g., PATE-GAN [JYS19] or MST [MMS21]) and (2) by designing a full pre-processing pipeline that embeds privacy-preserving calls to the server (e.g., through a privacy-preserving computation framework such as Ektelo [Zha+18]).

The massive access mode must also authenticate users in order to monitor the overall privacy guarantees satisfied for each user based on his disclosures log and on the composability properties of the privacy-preserving data publishing schemes used.

As a result, the data is protected using authentication and strong privacy definitions as presented in Section 3.2.2. Examples of applications of differential privacy to NLP models include [FDM19], which adds noise to word-frequency-matrix to achieve differential privacy, or [WK18], which samples the dictionary of the model using the differentially-private exponential mechanism [DR14]. The aim of these two approaches is to protect against authorship attribution.

[FDM19] uses a relaxation of differential privacy, d_χ-privacy [Alv+18] which allows the authors to consider a distance between documents computed using word embeddings, rather than the row-based distance presented in Definition 1. The objective is to modify BoW representation of documents "similar in topics" remain "similar to each other" (w.r.t. the metrics defined on word embeddings), irrespective of authorship. In practice, this is achieved by drawing BoW where the probability of each word being associated to a document is distributed according to a Laplace probability density function.

The goal of [WK18] is to derive a differentially private synthetic feature vectors, keeping the theme of each document while preventing authorship attribution. Feature vectors map a set of words (the dictionary) to probabilities of the word appearing in each document. The main idea of the approach is to sample the dictionary from a reference dictionary (e.g., using synonyms from WordNet's synsets) using the differentially private exponential mechanism.

In practice, the massive access mode can be plugged into the existing platforms that store massive number of legal documents and already support the precise access mode, such as CourtListener or CanLII.

Finally, another potential need is the annotation of documents, which is the addition to terms, sentences, paragraphs or documents of metadata such as syntax information (e.g., verb or noun), semantic, pragmatic (e.g., presupposition and implicature). This

step is crucial in NLP, and is usually done manually, for example through crowdsourcing. Crowdsourcing-specific approaches for privacy-preserving task processing [KBK14] require to split the task (i.e., annotation of a set of documents) between non-colluding workers (e.g., at the sentence level) before aggregating the result. Such approach is compatible with our architecture assuming the aggregation is done locally on the platform.

3.4.2. System overview

We now outline an abstract architecture for a privacy-preserving data publishing system for legal decisions. Our objective is not to provide exhaustive implementation guidelines, but rather to identify the key components that such an architecture should possess.

Figure 3.3 depicts the proposed architecture. The precise and massive access modes are both protected by the `Authentication` module. The `Authentication` module can be implemented by usual strong authentication techniques (e.g., for preventing impersonation attacks). Authentication is necessary for enforcing the access control policy through the `Access Control` module and for maintaining for each user his `Disclosure Log`. The log contains all the successful access requests performed by a user. It is required for verifying that the overall privacy guarantees are not breached, e.g., the rate limitation is not exceeded for the precise access mode, or the composition of the privacy-preserving data publishing schemes, formalized in the `Composability Properties`, does not exceed the disclosure allowed. Finally, the `Privacy Parameters` contain the overall privacy guarantees that must always hold, defined by the administrator (e.g., rate limit or higher bound on the ϵ differential privacy parameter). The user may additionally be allowed to tune the privacy parameters input by a privacy-preserving data publishing scheme (e.g., the fraction spent in the higher bound on the ϵ differential privacy parameter) provided it does not jeopardize the overall privacy guarantees.

3.5. Conclusion

In this chapter, we analyzed the needs for publishing legal data and the limitations of rule-based redaction (i.e., the current approach) for fulfilling them successfully. We proposed to discard any one-size-fits-all approach and outlined a straw man architecture balancing the utility and privacy requirements by distinguishing the traditional, one-to-one, use of legal data from the modern, massive, use of legal data by legal technologies. This step back from current practice allowed us to–in some way–avoid the tension between social and technical requirements.

Although our proposition could easily be implemented on current platforms, components of our framework pose new research questions. For example, reasoning with composability properties from different privacy models such as presented in Section 2.1.2 is a challenge. Another issue, discussed in Section 3.2.2 is the difficulty to evaluate empirical properties of approaches such as pseudonymization. Such a question could be answered by following the approach we describe in the following chapter.

4. Empirical privacy evaluation

This chapter is adapted from [ABG23c], and from the first edition of the SNAKE challenge [ABG23b] organized during Summer 2023.

4.1. SNAKE framework

Competitions and challenges are commonly used both in the machine learning community–for boosting the design and development of practical and efficient solutions to hard or new problems–and in the security community–for training purposes or for the evaluation of existing infrastructures. In contrast, in the privacy community, there is not a long tradition of holding such challenges. However, in recent years several competitions focusing on data sanitization algorithms (also called *data anonymization algorithms* or *privacy-preserving data publishing algorithms*) have been launched [Bou+20; Jor+20; Mur+23; Rid+21].

Some of them, like the *2018 Differential Privacy NIST Challenge* [Rid+21], have focused primarily on the defense aspect. More precisely, their main requirements were that the proposed algorithms have to (1) meet formal guarantees such as differential privacy, (2) achieve high utility levels on real-life cases, and (3) are efficient enough for being run on today's off-the-shelf computer systems. Others, like the *Hide-and-Seek challenge* [Jor+20], the INSAnonym competition [Bou+20] or the PWSCUP [Mur+23] additionally consider attacks on the sanitized datasets generated by the participants. More precisely, these competitions were generally composed of two phases, in which the first one is dedicated to the design of sanitization algorithms (usually focus on a particular type of data and use case) while the second one usually consists in attacking the data sanitized with the algorithms developed during the first phase.

While the sanitization phases of past challenges have been successful, in the sense that it has provided insights on the privacy/utility trade-offs, lead to implementations from state-of-the-art sanitization algorithms or even novel algorithms, the outcomes of the attack phases were usually more mitigated. For example, the organizers of the Hide-and-Seek challenge have reported attack results equivalent to random guesses [Jor+20]. We believe that one of the main reason for this is that in order to be successful, the attack phase must allow participants (1) to dedicate sufficient time to the design, implementation and testing of their attack strategies and (2) to focus on a few algorithms.

In this chapter, we propose a series of privacy challenges called SNAKE specifically tailored to attacks against sanitization algorithms. More precisely, the general structure of the SNAKE challenge provides the following salient features: it (1) enables to concentrate the energy and expertise of participants against a small set of carefully chosen

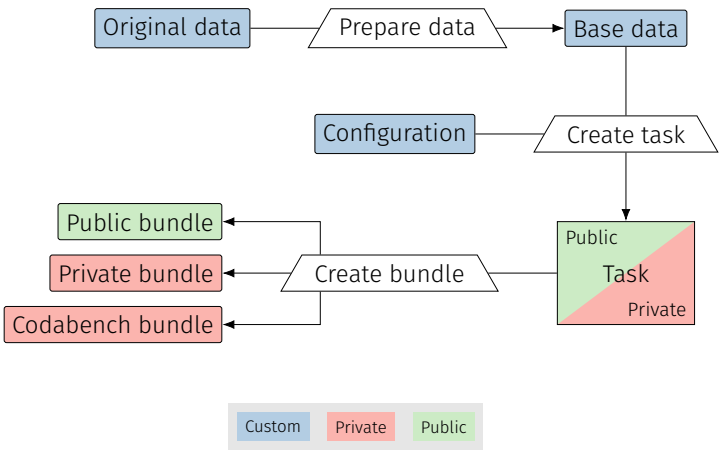


Figure 4.1.: High level view of the workflow. Rhombuses represent rules (i.e., programs) and rectangles data (i.e., files).

state-of-the-art sanitization algorithms, (2) enables to give sufficient time to participants to design and test carefully their attacks, (3) enables the exploration of a wide range of possible adversary models and background knowledge and (4) complements nicely the current sanitization competitions that focus on the sanitization part, thus resulting in a complete defense-attack pipeline. The first edition of the challenge focuses on the *Membership Inference Attack* setting [Hu+22; SOT22] (see Section 2.1) over tabular data synthetically generated while satisfying differential privacy [DR14]. The adversarial background knowledge consists in the record(s) of the target(s) of the attack. Further editions will showcase other inference attacks, adversarial background knowledge, privacy models and sanitization algorithms.

The abstractions defined in SNAKE consist in a workflow and its functional specifications. Figure 4.1 describes the SNAKE workflow. The rule Prepare data creates the main dataset of the challenge, called Base data, from some existing source of data, called Original data. Each task of the challenge is created by Create task, as a set of Public and Private files, with the former shared to participants, and the former used organizers (and the platform) for evaluation. The third rule Create bundle packages all the competition files into (1) a bundle can be uploaded to CodaBench to instantiate the full competition, (2) a public bundle which is to be hosted online to share the data necessary for participants and (3) a private bundle that is to be shared after the challenge.

4.1.1. Technical Foundations

SNAKE leverages (1) the Snakemake workflow management system [KR18] for its inner working and (2) the CodaBench competition platform [Xu+21] for all the features necessary for running a competition in real-life (e.g., user registration, submission management, leaderboard). Snakemake is a bioinformatics workflow engine which provides a domain specific language implemented as an extension to Python to describes pipelines. Workflows are specified as a directed acyclic graph of rules which transform input files

into target files, similar to *makefile* directives. SNAKE uses Snakemake to prepare all files required to define the challenge and outputs a "bundle" which can be uploaded to CodaBench to build a complete challenge.

4.1.2. Key properties

The design of the SNAKE framework adheres to the following key goals. They form the minimal set of properties that are both necessary to the successful organization of attack challenges targeting sanitization algorithms and general enough to adapt smoothly to the advances of the field.

Genericity The SNAKE workflow (see Figure 4.1) is high-level and *generic* enough to allow the design of most challenges in which participants attack the privacy properties of a sanitization algorithm. Indeed, any sanitization algorithm inputs prepared data and output sanitized data (e.g., a perturbed model), and any task consists in a challenge given to the participants (public part) and in the answer to the challenge (private part). In particular, the definition of a task as a pair of public and private data allows the framework to support a broad range of attacks scenarios (e.g., varying the background knowledge, the parameters, the kind of sanitization algorithm attacked).

Extensibility The SNAKE framework provides the two following levels of **extensibility**. First, it offers a simplistic challenge overview, which allows for further specification. For example, considering multiple attacks or allowing online attack evaluation is possible without deviating from the framework. Second, steps of the workflow are organized as Snakemake `rules` which (1) allow for generic input, output, script, execution environment, and parameters; and (2) can be broken down into sub-rules if the need arises.

Usability By leveraging the Snakemake workflow management, the SNAKE framework does not require strong development skills to be used and can be run on commodity hardware. It can thus be used easily by participants during a competition for building a complete local competition environment allowing local executions of the sanitization algorithms and of the participants' attacks. On the organizer side, existing challenges can easily be framed within SNAKE. Note that using SNAKE does not generate any computational overhead compared to a SNAKE solution. Moreover, the SNAKE framework outputs a package compliant with the CodaBench platform, which allows for user-friendly participation (documentation, live ranking, automated scoring using a Python script).

Reproducibility SNAKE allows a challenge to be reproduced in full, thanks to the functionalities offered by Snakemake. For example, inputs, outputs, execution environments, parameters and scripts are tracked through hashing. Random seeds used by Create task are saved in the private part of the related task. They can be kept by the organizers and disclosed to participants after the competition. The genericity and reproducibility

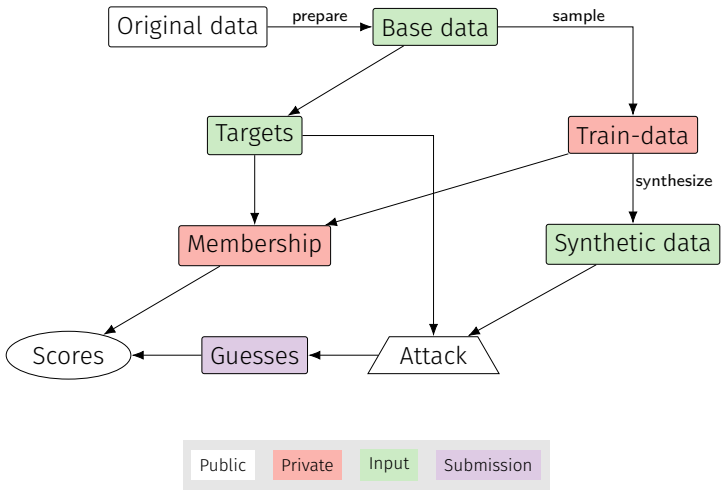


Figure 4.2.: High level view of the workflow implementation for the first edition. Rectangles represent data with rules as edge labels. The *Attack* rhombus is the (local) attack algorithm created by the participant. The *Score* ellipse is computed on CodaBench.

properties allow the framework to support different *threat models* over both participants and organizers. For example, our framework is designed to allow participants to verify locally that their scores, given by the organizer, are indeed correct. This can simply be done by the participant through a local evaluation of the submission performed after the distribution of the private part of the tasks. Preventing any tampering of the private part of tasks can be done, e.g., by storing their hash in the public bundle and by letting participants check that they match with the private parts distributed by the organizers after the competition.

4.2. SNAKE1: Membership inference attack against differentially-private synthetic data generators

SNAKE_1 is the first challenge based on the SNAKE framework. It focuses on *membership inference attacks* (e.g., [HHB19; Hye+22; Hu+21]) over differentially private synthetic data generation algorithms (e.g., [MMS21]), is co-located with APVP 2023 and takes place all along the Summer 2023.

Figure 4.2 describes the workflow of the challenge for a single task. First, the data used by the challenge is *prepared*–as described afterwards. Each execution of a sanitization algorithm takes as input a private dataset consisting of random *samples* from the base dataset. The attack used as input the *synthetic* data, a set of *targets* for which membership is to be guessed, the *base data* and the parameters of the sanitization algorithm. The participants then perform their attack locally[1] and submit their guesses

[1]Participants can use the framework to generate their own versions of the tasks locally and therefore

to the CodaBench space of SNAKE_1. The CodaBench platform scores the tasks by comparing the submission with the (private) ground-truth.

4.2.1. Attacks algorithm

Each team has to design an attack algorithm as follows:

Input The following information is provided to the attack algorithms:

- The synthetic dataset generated by an execution of the targeted sanitization algorithm over a private dataset;
- The targets to attack;
- The base dataset from which the private dataset is sampled;
- The parameters of the execution of the sanitization algorithm attacked.

Output The output of the attack algorithm is a real number in $[0, 1]$ indicating the predicted probability of each target being within the private dataset or not.

We detail below the computation of the private datasets, the targets, the sanitization algorithms as well as the attack success measure.

Base dataset and private datasets

SNAKE_1 makes use of the publicly available *EPI CPS Basic Monthly* data provided by the Economic Policy Institute [Eco23]. The CPS dataset is divided in years in which a yearly dataset contains more than 10^6 records and 125 columns. A record contains information about a single individual in a household. Our base dataset is built by using a pre-processed sample of columns and rows from the original dataset. Full details is available in the competition documentation. From this base dataset, we generate one private dataset for each parameterized sanitization algorithm attacked.

Targets and background knowledge

In SNAKE_1, any household that contains at least 5 individuals might be a target, with the target consisting of the full set of records of the household. SNAKE_1 considers the following background knowledge about each target. The adversary knows (1) the exact records of the household targeted, and (2) the full base dataset[2]. The adversary is also given the information about the sanitization algorithm targeted as well as the parameters used for the executions and has access to its implementation. However, the randomness generated internally during the execution of the algorithm is unknown to the adversary.

can evaluate their submission on their side before submission.

[2]It gives teams the knowledge of the population distribution commonly assumed in membership inference attacks.

Sanitization algorithms under attacks

The sanitization algorithms under attack during SNAKE_1 are differentially-private synthetic data generation algorithms. More precisely, we have selected a set of algorithms according to the following two criteria: technical soundness assessed by a rigorous peer-selection process (e.g., published at top-tier conferences or winner of a dedicated competition) and available open-source implementation. In particular, we have used the implementation available in the Reprosyn package [Mol23]. Except for parameters related to differential privacy, we use the default values set in their implementations. The PrivBayes algorithm [Zha+17] generates synthetic data by capturing the underlying distribution of the private data through a specific Bayesian network. The MST algorithm [MMS21] is a generalization of the NIST-MST algorithm, which has won the 2018 NIST Differential Privacy Synthetic Data challenge. It generates synthetic data by perturbing the marginals that capture the data distribution through the Gaussian mechanism and by post-processing them through the Private-PGM algorithm [MSM19]. The PATE-GAN algorithm [JYS19] is an extension of *generative adversarial networks* [Goo+20] based on the *private aggregation of teacher ensembles* framework [Pap+17].

Success measure

The success of a team is computed by first measuring the successes of its attack on each parameterized algorithm attacked (e.g., MST parameterized by $\epsilon = 1.0$, $\delta = 10^{-5}$) and second by aggregating the success measures in a single final score. The success of a given attack for a given parameterized algorithm is evaluated based on the *membership advantage measure* [Yeo+18] (see 6). The number of targets must be at the same time large enough for obtaining a sufficiently stable estimation and small enough for being practical (e.g., $r = 100$).

Execution environment

The design, coding, and executions of attacks are performed locally by each team with its own resources. As a result, there is no restriction on the computing environment used to develop the attacks and the choice of the programming language and the available resources are unconstrained. Note however that we provide to participants the SNAKE_1 Python environment.

4.2.2. Preliminary results

The initial ending of the challenge was planned to September 1st, 2023, which we then moved to October 1st. At that time, the challenge had 21 participants, as individuals or teams. Due to technical issues with the platform we had to postpone the ending by another month, to November 1st, which prevents us for presenting the detailed results of the challenge.

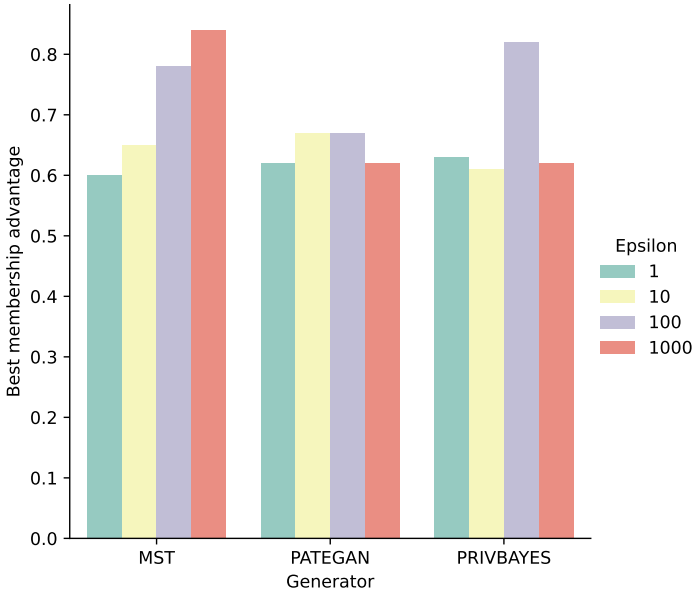


Figure 4.3.: Best membership advantage scores on October 5$^{\text{th}}$, 2023

As a first feedback, we note that the best scores are not attained for the highest values of ϵ, as could have been expected. Figure 4.3 shows the best score among all submissions for each task. The results for MST are as expected, with scores increasing with a larger budget. The best results for PATEGAN are close to each other. A first hypothesis might be that the learning error is higher than the impact of the noise added by differential privacy. Lastly, the values for PrivBayes are the most surprising: three of the tasks have similar scores, with the exception of $\epsilon = 100$ which exhibits a particularly high membership advantage.

4.3. Future editions

The SNAKE framework is dedicated to facilitating the organization of challenges stress-testing sanitization algorithms. Overall, we hope that the SNAKE challenges can help gain understanding of the empirical privacy guarantees of sanitization algorithms and of their parameters and pave the way to a greater democratization of sanitization algorithms providing strong privacy guarantees. We envision SNAKE as a series of challenges, with future editions focusing on different datasets, privacy mechanisms and frameworks, attack models, etc.

Though first envisioned as a proof of concept, we created a toy competition named EKANS which could be the basis for a future edition. It is thought as the "reverse" of SNAKE$_1$, in that its objective is to propose a sample method for targets that are the easiest to attack. EKANS asks participants to submit *Python code* in charge of sampling "good" targets for membership inference attacks, which is executed by the CodaBench platform before performing a basic attack on the sampled targets using the TAPAS

package [Hou+22].

Finally, as a future work, and considering the technical issues that we faced, we are considering building a platform customized to the SNAKE framework. This platform could benefit from the generic architecture and automated production of tasks to, in addition, deploy an API and database tailored to the challenge as described in for Snakemake.

5. Simulating long-term discrimination

In this chapter we focus on the fairness issues presented in Section 1, in which we highlighted the importance of considering the long-term impact of (un)fairness, and associated concepts that are often discarded in formal formulations of fairness. We start by advocating for the need of using a simulation for studying long-term evolution of (un)fairness mechanisms. Then, we present the main components of the model we adopt [Rea+18], followed by a overview of the efforts need to undertook its reproduction.

5.1. The case for a simulation

As a first step towards the addition of long-term concepts in fairness, we need to be able to observe and identify effects of fairness effects in a long-term context. To achieve this goal, we adapted an agent-based simulation originating from the domain of educational policy, which, we hope, ground our work in reality–although highly simplified.

The use of a simulation is motivated by multiple factors. First and foremost, experimenting with fairness–and therefore unfairness–in high-stakes contexts cannot be done in an ethical fashion. This, the ability to draw observations without impacting individuals is obviously needed in this context. Second, the long-term effects in which we are interested requires long-term observation, which itself poses challenges. In particular, following individuals on the long-term is expansive and difficult [GG07]. Additionally, with observations being done over a long time period, more factors have to be taken in account since the number of decisions taken or received by an individual will increase. This is an effect of the *curse of dimensionality*, and directly linked to the argument towards the consideration of a snowball effect. Indeed, if decisions accumulate with time, the presence of unforeseen or uncontrolled decisions will make observations noisy.

Studying the long-term impact of a chain of decision, as an observation and without intervention, still requires access to a dataset following individuals during their life. Accuracy of measurements would require many individuals while meaningful observations would require collecting data over a long duration. Additionally, experimenting using machine learning models requires a large training set. To the best of our knowledge, no such dataset exists at the best of our knowledge. For example, the United States Census reports a high number of features on many individuals [Abo18], but longitudinal linkage is limited [Hel+22].

As a result, observing the *impact* of the decision is not possible without simulation–or would require taking high-stakes decisions with all models to be experimented with in real life, which would be unethical and impractical. A simulation has the benefit of providing any number of decisions over any period of time. Additionally, the bias

introduced by the simulation is known and controlled. On the other hand, the realistic aspect of the simulation is a concern. We opted for the model described in [Rea+18] that studies the impact of affirmative action–which refers to the set of practices that seek to include particular groups that were discriminated against–in college admissions using a simulation. This is motivated by several factors. First, this model comes from the education policy community and is used to draw conclusions on a real-life process. In particular, the authors note that "although [their] model falls short of being completely realistic, it captures important, dynamic features of the application/admissions/enrollment processes that enable the investigation of the ways that affirmative action might affect enrollments." In addition, parameters are chosen to fit observation of the *Education Longitudinal Study of 2002* [Nat02]. Second, the model is described extensively. Although we faced several difficulties regarding the implementation–typical of the issue of reproduction in science, particularly in an interdisciplinary context–the paper makes extensive efforts to describe the model in details. Finally, we required a simulation that considers bias. For that matter, we searched for models in social sciences that cited "race". This potentially fits any simulation using sensitive attributes (such as gender or race), but the case study of [Rea+18] considers the issue of discrimination itself, doing so through the correlated attributes of race and socioeconomic status.

5.2. College-student model

This section provides an overview of the model described by [Rea+18]. A more technical description is available in Appendix A. We first introduce the two types of agents that interact in this simulation: colleges and students. Then we go over the three successive decisions that are taken at each iteration:

1. students apply to a subset of colleges depending on their own preferences and probability of being accepted;
2. colleges admit the top applying students;
3. students enroll into the best college they are admitted into.

Figure 5.1 is a complete view of the causal relationships between variables, which are described in Table 5.1. We will present relevant subsets of this figure along the section when discussing specific parts of the model. Specifically, Figures 5.2, 5.2, and 5.4 focus respectively on the application, admission and enrollment decisions.

5.2.1. Colleges and students

The model is composed of $J = 40$ colleges and $N = 10\,000$ students.

Colleges are represented by their *quality* attribute, initially drawn from a normal distribution and later updated depending on the enrolled students, as represented by the orange arrows on Figure 5.1.

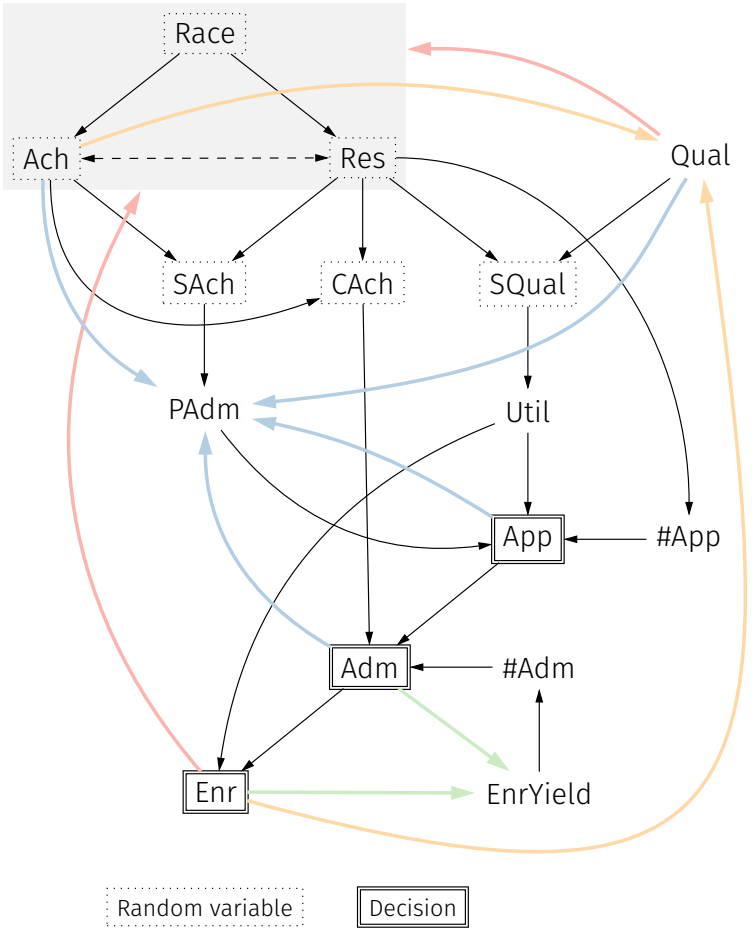


Figure 5.1.: Simplified causal model. Colored arrows represent causal relationship on future states.

Variable	Description
Race	Student's quality
Ach	Student's achievement
Res	Student's resources
Qual	College's quality
SAch	Student's perception of their own achievement
CAch	College's perception of a student's achievement
SQual	Student's perception of a college's quality
PAdm	Student's estimation of its probability of admission into a college
Util	Student's perception of the utility of a college
#App	Student's number of applications
#Adm	College's number of admissions
EnrYield	College's estimation of its enrollment yield
App	Student's application to a college
Adm	College's admission of a student
Enr	Student's enrollment into a college

Table 5.1.: Description of variables of the college-student model

Students are represented by three variables: race, resources and achievement. The *race* of each students is drawn from a non-uniform distribution over the categories Asian, Black, Hispanic, and White. Note that, following [Rea+18], the numerical majority is White, while Asian is the least represented race, and that "minority" corresponds to Black and Hispanic students. *Resources* and *achievement* are sampled from race-specific bivariate normal distributions with non-zero correlation. The students' mixture distribution is:

$$\begin{aligned} Race &\sim \text{Categorical}(\theta_{Race}) \\ Res, Ach \mid Race &= Normal_2(\mu_z, \Sigma_z) \end{aligned} \tag{5.1}$$

5.2.2. Application

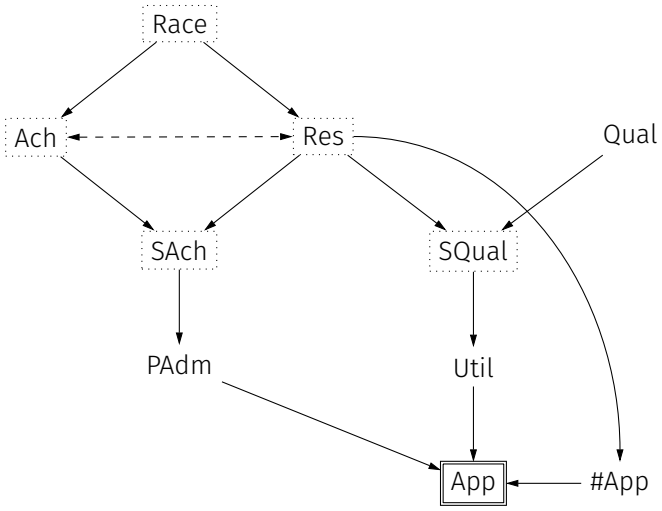


Figure 5.2.: Simplified causal model of the application decision.

During the first stage, illustrated on Figure 5.2, students select a subset–called portfolio– of colleges to apply to. Students observe colleges' quality with some amount of uncertainty representing "imperfect information and idiosyncratic preferences". This error depends on students' resources as a way to model high-resources families having better information about college quality. Based on their noisy observations of their own achievement and college quality, students estimate their probabilities of admission into each college using a logistic model fitted on admission patterns over the past five years, represented by blue arrows on Figure 5.1. Each student applies to a subset of colleges that maximize their overall expected utility. More precisely, the expected utility of an application portfolio is computed as:

$$\begin{aligned} EU(\emptyset) &= 0 \\ EU(c_1, c_2, \ldots, c_n) &= Util(c_1) \cdot PAdm(c_i) + (1 - PAdm(c_1) \cdot EU(c_1) \end{aligned} \tag{5.2}$$

with $c_1, c_2, \ldots, c_n$ ordered by descending utility.

The model assumes that all students are rational, utility-maximizing agents. The variability in student application portfolios originates from the imperfect observation of variables.

5.2.3. Admission

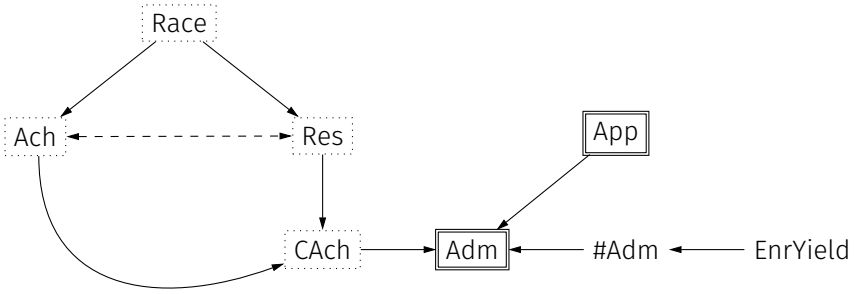


Figure 5.3.: Simplified causal model of the admission decision.

The second decision phase, illustrated by Figure 5.3, consists in the admission of students by colleges. Colleges observe the achievement of applicants with some noise, similar to the observation of quality by students and serves the same purpose of modeling uncertainty and idiosyncratic preferences. Colleges rank applicants according to this variable and select the top subset of applicants, with the objective of enrolling 150 students. The number of admitted students is computed by each college from its previous enrollment yield, represented by green arrows on Figure 5.1.

5.2.4. Enrollment

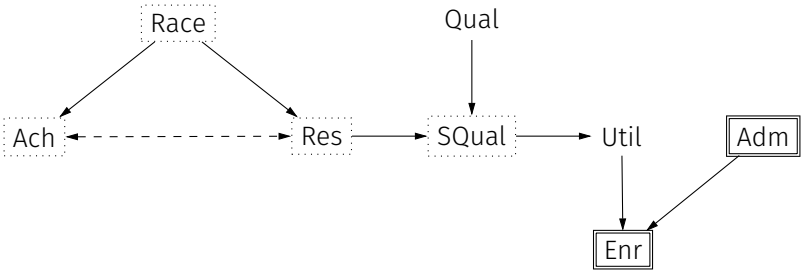


Figure 5.4.: Simplified causal model of the enrollment decision.

The final decision, illustrated by Figure 5.4, is straightforward. Students enroll in the college with the highest estimated utility to which they were admitted.

5.2.5. Fairness interventions

Three fairness mechanisms are considered by [Rea+18]: two affirmative action bonuses and one "targeted recruitment". The methods are introduced after the 15^{th} iteration by a subset of colleges.

Affirmative action can be based on race or socioeconomic status (SES). During the observation of students' achievement by colleges, an additional weight is added. In the case of race-based affirmative action, the added weight is constant added to minority students, that is Black and Hispanic students as introduced in Section 5.2.1. In the case of SES-based affirmative action, a weight is applied in accordance with each student's *resources*, such as $150 \times Res$. The actual weight varies between experiments, and is a source of difficulties regarding the reproduction which we discuss later in Section 5.4.

Targeted recruitment is the targeting of underrepresented racial minority students. It is represented as an increase in the observed utility of attending a college by students. Its implementation in reality would require colleges to take action towards students through targeted communication.

5.3. Reproduction

5.3.1. Method

Although the model is described in details, missing detailsinformation and inconsistencies hindered its reproduction [ABG23a].

We have mostly followed the detailed description of the model given in [Rea+18, Appendix C]. We have also referred to [Rea+16], which is often cited by the authors as an inspiration for this work. Code for this latter model is available in the (proprietary) Stata programming language [Rea+14]. We have read this code for clarification, but we were not able to run it[1]. We have contacted the first two authors of [Rea+18] (Sean F. Reardon and Rachel Baker) in May 2021 but got no answer as of today.

For our implementation, we have used the Python programming language, with most of the processing done using NumPy [WCV11]. We have used scikit-learn [KK16] for learning the logistic regressionWe have experimented with the Xarray library [HH17], which provides labeled multidimensional arrays. Our motivation was to deduplicate the join between students and colleges. However, we have achieved no storage or time improvement by doing so.

In addition to the replication effort, we focused on execution speed to be able to use the model as a synthetic data generator in other works. For this purpose, random data for all future iterations is drawn at initialization, and we took extensive care to use *vectorized operations* provided by the Numpy framework. We do not provide an extensive benchmark but still enjoy more than 5 iterations per seconds on off the shelf hardware.

[1] Executing this model seems to require Stata version 11, but the dependency "`college sorting napps.do`" is not distributed in the archive.

5.4. Difficulties

In this section, we present the main issues encountered with respect to the description of the model in the original paper [Rea+18]. Hereafter, inconsistent descriptions are quoted, but we refer the reader to the full paper for more context. Afterwards, we describe the interpretation that we have followed for the results presented in this paper.

Socioeconomic status (SES)-based affirmative action.

- "It is during the calculation of A^{**}_{cs} that colleges with an affirmative action policy apply additional weight to a student's perceived admissions desirability in accordance with that policy. This additional weight is captured by the term $T_c \times [G \times (Black_s \mid Hispanic_s) + H \times resources_s]$. In this term, T_c indicates whether a college has an affirmative action policy, [...] H is the size of the weight given to students under SES-based affirmative action policies, which is applied *linearly in accordance with the student's resources*, $resources_s$." [Rea+18, Section 3]

- "SES-affirmative action (corresponding to an increase in admissions consideration of 0, 50, 100 and 150 *achievement points for each decrease of one standard deviation* in applicants' resources);" [Rea+18, Section 4].

Using $H \times resources_s$ as in the first quote would advantage high-income students, which is the opposite of the objective. Moreover, the weight would not follow the description of the next two quotes, which involve the standard deviation. In our implementation, a weight $H \times \max(-\text{zscore}(resources_s), 0)$ is added to the score of low-income students, with zscore being the standard score[2] with respect to $resources$, and H the size of the weight (e.g., 0, 50, 100 and 150). We keep only negative values of the standard score (see Figure B.2 for a version without truncation) to impact only low-income students ("for each decrease of one standard deviation"). Note that while our interpretation is not totally satisfactory (see Section 5.5), the formula of the first quote (see Figure B.1) was equally unsuccessful.

Quality and own achievement reliability.

- "Quality reliability and own achievement reliability *bounded by minimum values of 0.5 and maximum values of 0.9.*" [Rea+18, Table 1]

- "the reliability of student perceptions of their own achievement, is a function of student resources and *bounded between 0.5 and 0.9*" [Rea+18, Appendix C]

- "the reliability of student perceptions of college quality, is a function of student resources and *bounded between 0.5 and 0.7*" [Rea+18, Appendix C]

We assume that 0.7 is a typo and clip the values between 0.5 and 0.9. However, the impact is quite negligible as seed by comparing Figures 5.7 and B.3.

[2]The zscore variable is the number of standard deviations by which the value of a raw score is above or below the mean value of what is being observed or measured

Race-targeted recruiting. "Colleges' binary recruitment statuses (S_c)–which had previously all been 0–are set based on model parameters that determine which colleges will use recruitment [...] Utility is then calculated using model-specific recruitment magnitude values (L): $U^*_{cs} = a_s + b_s \times Q^*_{cs} + R_{sc}$. " [Rea+18, Appendix C]

There is no use of S_c in the document apart from the definition, and R_{sc} is also not defined. We assume that it is a typo and that S_c is to be understood as R_{sc}. The recruitment magnitude value L is not used (but is an experimental parameter later). No specific definition of race-targeted recruitment is given but from "Recruitment efforts work in part by making students aware of specific colleges and by making these colleges seem more appealing to prospective students through additional, targeted contact with those students." [Rea+18, Section 3] our understanding is that targeted groups have an increased chance of applying to colleges using race-targeted recruitment. However, the race attribute is not used in conjunction with L.

We assume that the race-targeted recruitment weight is added to the perceived utility of attending colleges for minority students, such that

$$U^*_{cs} = a_s + b_s \times Q^*_{cs} + R \times (Black_s \mid Hispanic_s) \times T'_c$$

in which T'_c indicates whether a college uses race-targeted recruitment and R is the recruitment weight.

Quality update window. It is unclear whether the quality of colleges depends only on the previous year or on the past five years:

- "Colleges' quality values (Q_c) are *updated based on the incoming class of enrolled students before the next year's cohort* of students begins the application process" [Rea+18, Appendix C].
- "College quality is *calculated as the five-year running average* of enrolled student caliber." [Rea+18, Figure C5]

The later point is given in the caption of a figure, whereas the first is part of the model description. Using only the previous year to update colleges' quality gives satisfying results (see Figure 5.8). Thus, we assume that the five-year running average is not used for the model but only to smooth the plots.

Initial college quality. "Initial college quality (Q) is normally distributed" [Rea+18, Appendix C], but the values of the mean and variance are not given. We relied on the values from [Rea+16]: $Q \sim \mathcal{N}(1070, 130^2)$.

Admission probability estimation. After the fifth iteration, each iteration requires fitting a logistic model on submitted application over the past five years. This allows students to estimate their probability of admission into each college, which is required for them to select a subset to apply to.

Description	Domain
Number of years	30, 50
Number of top colleges implementing policy	0, 4
Race-based affirmative action weight	0, 150, 300, 260
SES-based affirmative action weight	0, 50, 75, 100, 150
Race-based recruiting weight	0, 25, 50, 100

Table 5.2.: Experimental parameters

Apart from a significant speed increase, we observe no difference when fitting on the previous year only (and using the previously fitted model's parameters to initialize the new fit). However, our code still uses the past five years.

Weights of affirmative actions and targeted recruitment. The actual mapping between weights and their label changes between "*moderate* SES-based affirmative action and moderate race-based recruitment, which corresponds to a weight of 100 and 75, respectively." [Rea+18, Figure C3] and "*strong* SES-based affirmative action and strong race-based recruitment, which corresponds to a weight of 100 and 75, respectively." [Rea+18, Figure C3]

5.5. Experiments

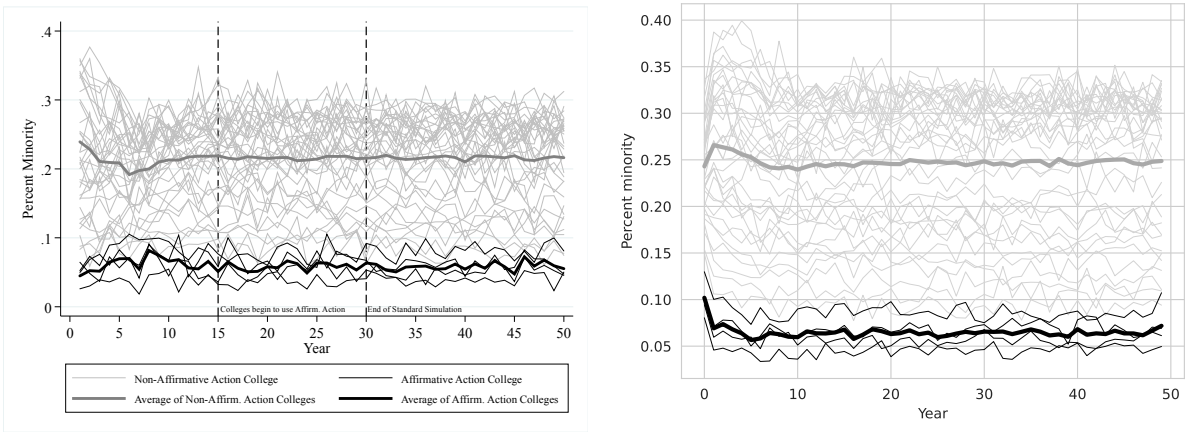
This section contains original figures from [Rea+17] along with our replications, followed by discussions of each result. To ease readability, we force the positioning of figures, which is therefore not optimal. We apologize to the reader for the resulting empty spaces. Furthermore, please note that to ease comparison we attempt–to the best of our ability–to reuse the style of the original figures, even when not ideal. We also apologize for the gray color palette. Original figures in the published version of the paper [Rea+18] are colored. However, because of the high costs of the reproduction rights, we chose to use the black and white versions of the preprint.

Our objective was to replicate the overall tendencies of the model. We compare below our replicated results to the original results.

Figures 5.5 to 5.14 present side by side examples of reproduction, with plots taken from the preprint version [Rea+17] of the original paper [Rea+18] (due to copyright issues) along with our replication results. We attempt to replicate a representative subset of the plots, covering the gist of the model: impact of race-based affirmative action (Figures 5.6, 5.13, and 5.14) and impact of SES-based affirmative action and race-based recruitment (Figures 5.7 to 5.12). Additional replications are available in Appendix B. Table 5.2 describes the available parameters and the values used in the experiments.

Figures 5.11 to 5.14 present the impact of the different strategies on the final (years

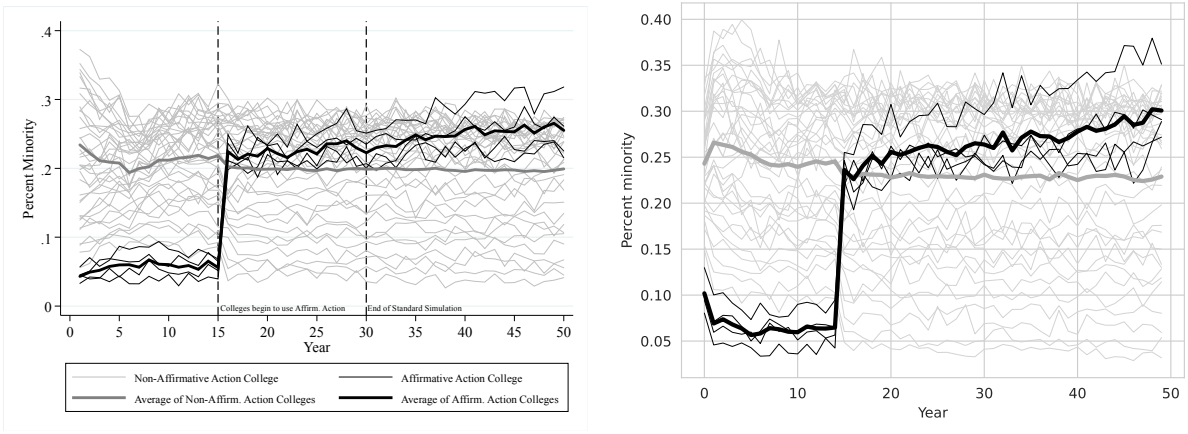
25 to 29) composition–racial and socioeconomic[3]–of enrolled students in colleges. Our replication effort involved experimenting with the full set of plots presented in the original paper. We also resorted to additional plots (not presented in the original paper or the current one) to assess the correlation between admission probability and academic achievement. Every experiment results from an average of 10 runs.



(a) Original [Rea+17, Fig. C1] (b) Replication

Figure 5.5.: Minority enrollment without using affirmative action or recruiting.

The results without using positive discrimination policy presented on Figure 5.5 are similar to the original paper. The first years appear to exhibit a reversed trend, which could be due to difference in the way the logistic regression is being fitted.

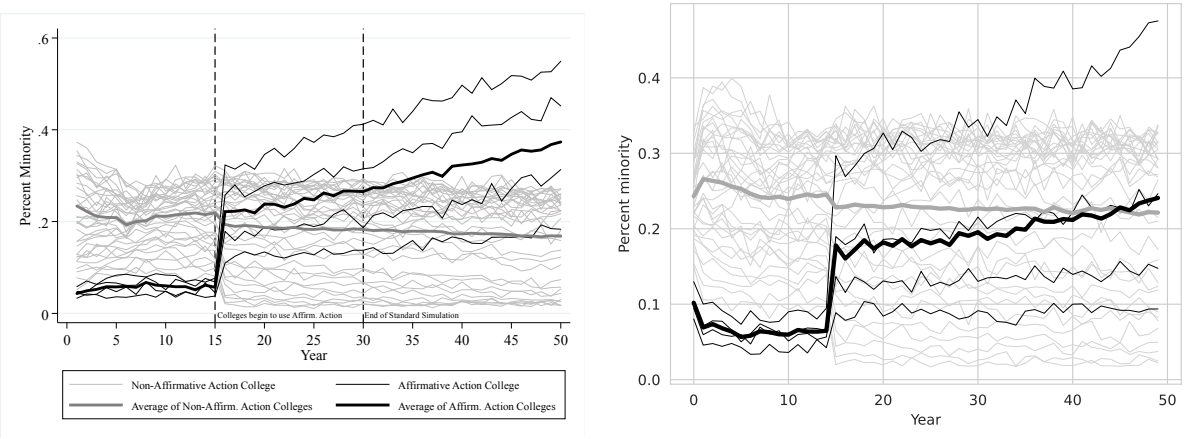


(a) Original [Rea+17, Fig. C2] (b) Replication

Figure 5.6.: Minority enrollment with the top four colleges using real-world race-based affirmative action (weight of 260).

[3]Our socioeconomic distribution is represented by the quintiles of the *full* students' resources distribution (not only the enrolled students) as doing otherwise results in incomparable categories.

Figure 5.6 shows that we replicated the main trends but still indicates an issue with race-based affirmative action: our replication leads to a greater impact on minority students.

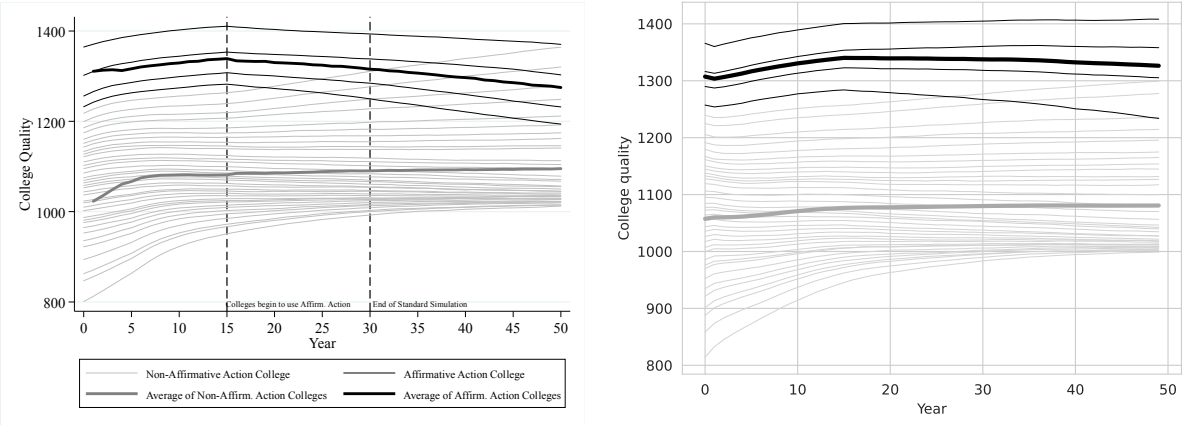


(a) Original [Rea+17, Fig. C4]

(b) Replication

Figure 5.7.: Minority enrollment with the top four colleges using strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100).

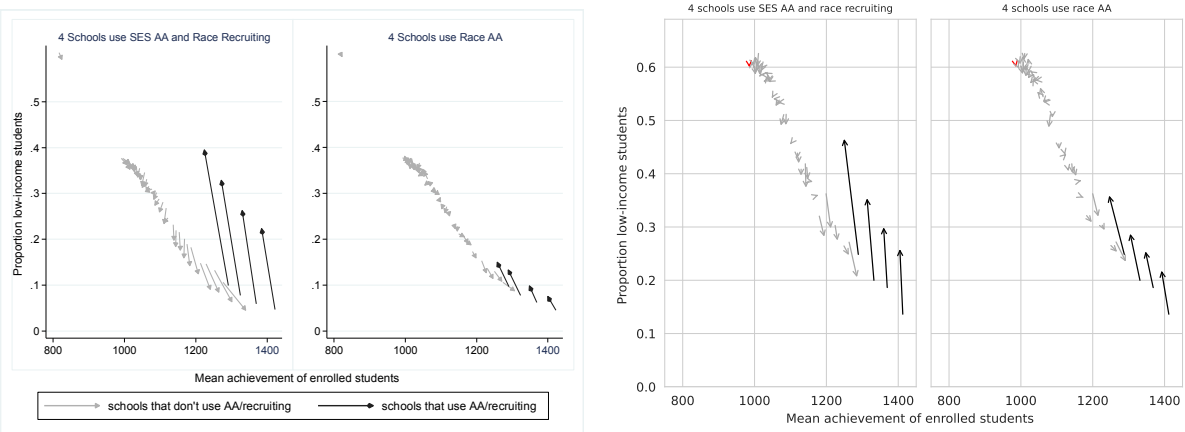
Looking at Figure 5.7, the combination of SES-based affirmative action and race-based recruitment is satisfactory in terms of range, but the average impact on minority students is lower than the original. This is also the case with Figure B.4 in Appendix B which considers the same parameters. However, race-based and SES-based affirmative actions use the same mechanism, which leads us to believe that race-based recruitment is successfully replicated with a potential issue for SES-based affirmative (as with Figure 5.6).



(a) Original [Rea+17, Fig. C5] (b) Replication

Figure 5.8.: College quality with the top four colleges using strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100).

The behavior of quality on Figure 5.8 is similar to the original paper, indicating the adequacy of our quality initialization. A strong claim in the original paper is the decline in quality of colleges using policies, which is not as clear in our replication as only two of the four active colleges suffer such decline. This might be attributed to the lower impact of our actions, as highlighted by the other figures.



(a) Original [Rea+17, Fig. D1] (b) Replication

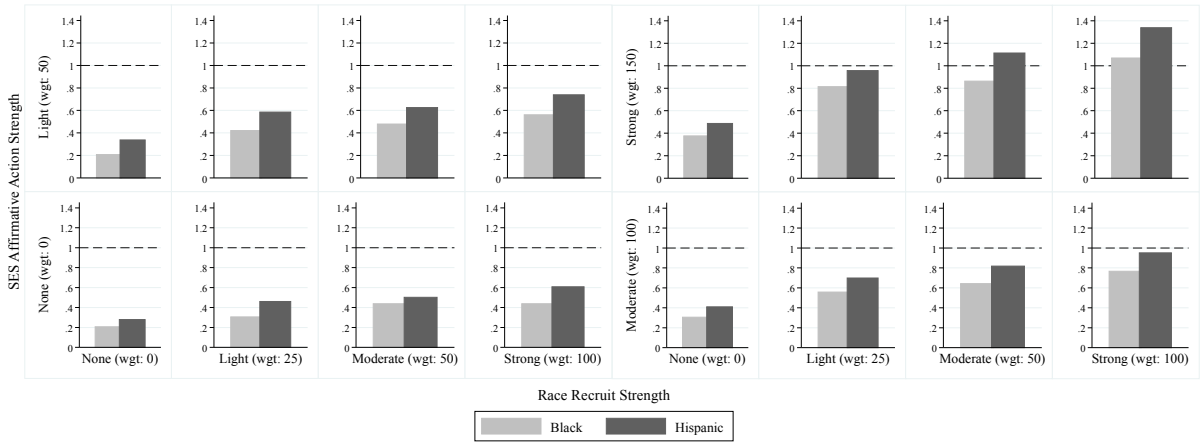
Figure 5.9.: Mean achievement and proportion low-income with top four colleges using (respectively not using) strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100) on the left sides, and real-world race-based affirmative action (weight of 260) on the right sides. Arrows start at year 14 with no affirmative action, and end at year 29.

Figure 5.9 suffers from the same issue of affirmative actions as Figure 5.6, but the general behavior is similar to the paper. Figures B.5 and B.6 in Appendix B present the

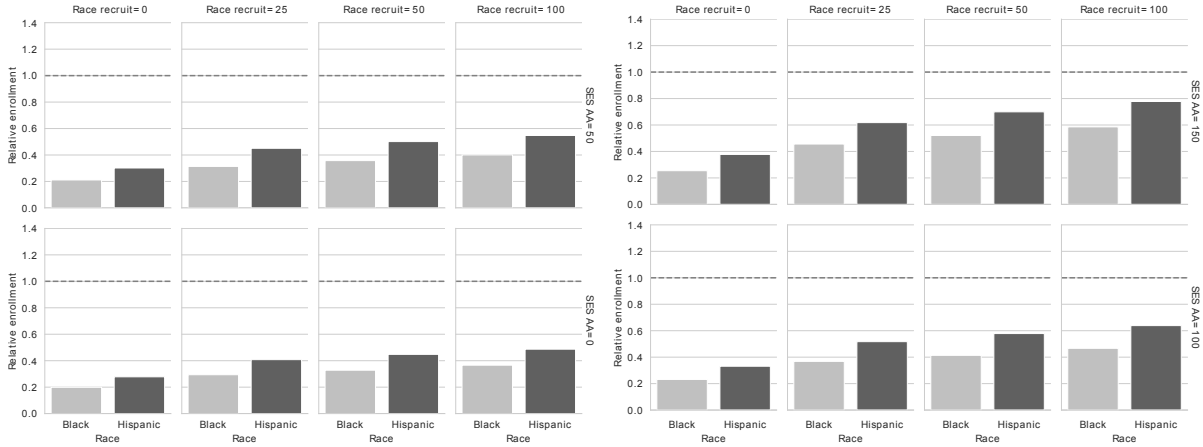
same plot for minority and low-income students with varying number of active colleges.

It is important to note that this figure corresponds to an average over 10 runs (as the original paper), which requires the colleges to be sorted by quality.

The additional left-most arrow covers students who do not enroll. However, it is unspecified however whether it captures students who do not enroll while having applied, being admitted or without condition. We use the latter (see Appendix B).



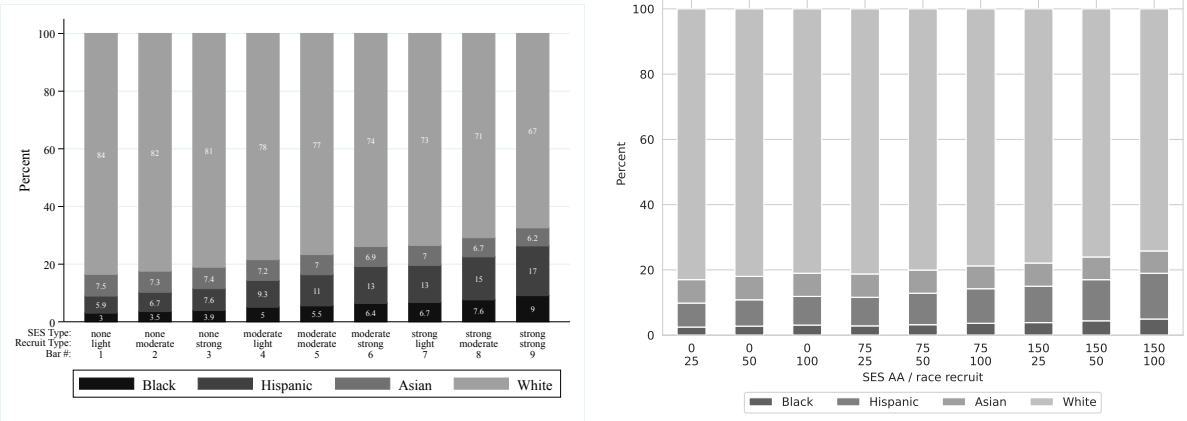
(a) Original [Rea+17, Fig. 2]



(b) Replication

Figure 5.10.: Black and Hispanic enrollment in colleges using SES-based affirmative action and race-based recruitment, as a share of estimated enrollment under race-based affirmative action (using estimated real-world affirmative action weight 260).

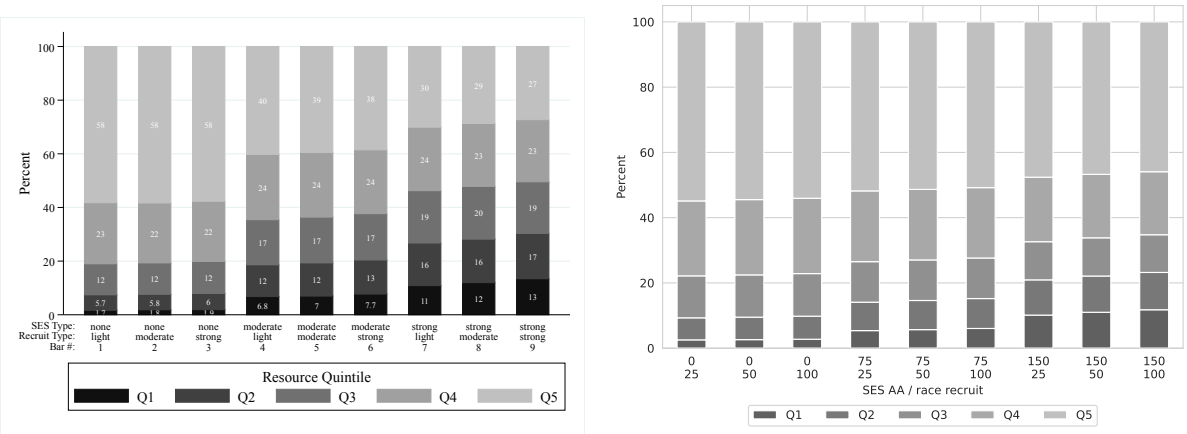
Figure 5.10 shows a successful replication of the general tendency of the model (i.e., the relative values) and a failure in replicating the absolute numbers, probably due to the issue related to the affirmative action observed above.



(a) Original [Rea+17, Fig. A2]

(b) Reproduction

Figure 5.11.: Racial composition of colleges using SES-based affirmative action and race-based recruitment, by affirmative action and recruitment weights.

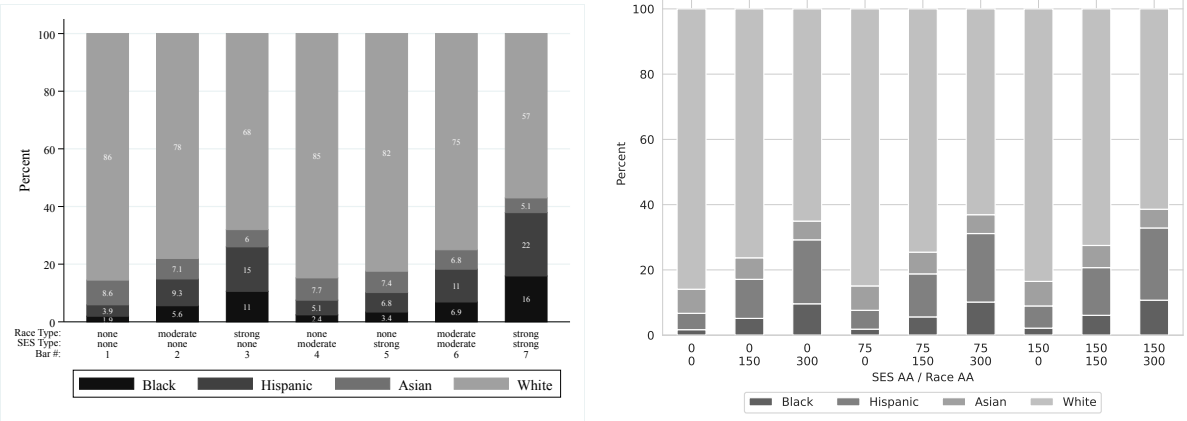


(a) Original [Rea+17, Fig. A3]

(b) Reproduction

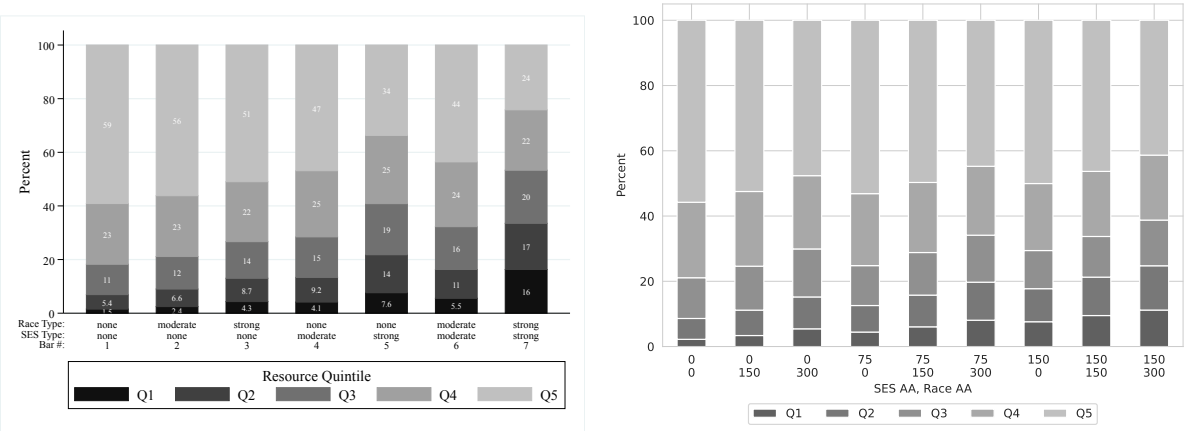
Figure 5.12.: Socioeconomic composition of colleges using SES-based affirmative action and race-based recruitment, by affirmative action and recruitment weights.

Figures 5.11 and 5.12 shows a successful replication regarding the racial impact of targeted recruiting and SES-based affirmative action strategies, but highlight an issue with the socioeconomic composition (Figure 5.12). Indeed, the impact of our SES-based affirmative action is higher than the original paper.



(a) Original [Rea+17, Fig. A4] (b) Reproduction: racial composition

Figure 5.13.: Racial composition of colleges using SES-based and racial affirmative actions, by affirmative action weights.



(a) Original [Rea+17, Fig. A5] (b) Reproduction

Figure 5.14.: Socioeconomic composition of colleges using SES-based and racial affirmative actions, by affirmative action weights.

Figures 5.13 and 5.14 lead to the same observation that SES-based affirmative action is not behaving as the original paper. However, race-based affirmative action is satisfactory. Note that the combinations (race weight = 150, SES weight = 150) and (race weight = 300, SES weight = 75) are not presented in the original paper.

Result Although we achieve mostly partial successes, the overall behavior is successfully reproduced, with discrepancies regarding the impact of positive discrimination policies. We attribute the latter to issues with our understanding of the paper, thus highlighting the difficulties of describing and understanding complex systems, and the need for open-access to code. The resulting implementation, while not producing results

entirely faithful to the ones from the original paper, can still be useful to simulate data grounded in real-world in the context of college admission.

5.6. Towards an unfairness warning system

This section presents the next step towards building an unfairness warning system based on the model presented in this chapter.

5.6.1. Long-term dynamic

To observe the impact of fairness concepts on the long-term, we need to model two concepts of long-term impact: the feedback loop and a "snowball effect". The first is already encompassed in the dependencies across states, since variables which represent the context of the decision are impacted by previous decisions, mainly quality of the colleges and probability of admission.

Snowball: decision accumulation

To model the snowball effect, we add a dynamic aspect to the students, so that innate attributes of individuals are impacted by previous states. Our original definition of the snowball effect considers decisions at each steps as features for the next. This would imply (1) an ever-growing accumulation of features and (2) a require several modifications to the decision process itself, whereas we try to keep the model as is. Instead of following this individual formulation, we could apply the accumulation on the groups: the weights and distribution of resources and achievement of racial groups are modified according to their respective enrollment, as represented by the red arrows on Figure 5.1. This formulation is equivalent to our original definition if we consider an aggregating step: rather than appending decision y to the features X, we set the new features as a function of X and y. A way to view our proposition is to consider students as children of the parent cohort, benefiting from their successes. In short, we can model the impact on students in two ways:

Representativity The weights of the race distribution can be modified to fit the enrollment success of each race. This impacts the representativity of each group, and models a unique cohort of students going to successive stages of selection.

Reproduction The race-based weights of achievement and resources can be penalized or advantaged depending on the enrollment rate of past cohorts, potentially taking into account the quality of colleges in which students where enrolled. This would correspond to future students benefiting or not from their parents academical background.

5.6.2. Interventions

The model allows for a broad kind of fairness intervention policies. First, it mixes different kinds of decisions: (1) a numerical estimation of the probability of admission used by students, (2) an optimization problem to build the application portfolio, (3) a top-k ranking for the admission of students, as well as finally a top-1 ranking for the final enrollment. Since each mechanism can independently use noisy observation or perfect observation, different combinations can be envisioned. In particular, the point (3) allows our model to be used to evaluate fair rankings [ZYS23], while the point (1) would allow tools such as [Bel+19] to be plugged in, providing multiple fairness-aware machine learning models.

Apart from algorithmic interventions, the way bias is added through observation of variables by students and colleges (e.g., through the SAch, CAch, and SQual variables), allows for more classical fairness policies to be studied. Affirmative action, in the form of a bonus weight added to minority groups, is already part of the model. Random noise can be used as a baseline by removing the correlation between observation noise and race.

5.6.3. Reference scenarios

In addition to providing a playground for multiple fairness-aware algorithms, the model allows for many different fairness assumptions. As discussed in Section 2.2, the particular state of reference used to evaluate fairness is crucial, and a matter of social consideration. Leveraging a simulation is particularly interesting here since we can compare the impact of using a particular reference. Figure 5.15 illustrates such options with red arrows. Fair transitions of the alternate chain can be achieved by accessing the pre-bias students variables (i.e., "innate") resources and achievement, which are available as random variables before mixing. When evaluating fairness at state S_i, several references can be used:

Previous using the previous state is the typical choice of many fairness papers, with the data used to "train" the current state S_i is used as reference;

Past similarly, an arbitrary state in the past S_j can be used, increasing the window of observation;

Initial although generally undefined in real-world scenarios, we can also use the initial state of the simulation S_0;

Alternative more interesting, we can use for reference an alternative state $S_{i''}$ which is the result of using an alternative transition, for example one implementing a specific fairness intervention, as defined above;

Parallel finally, as an option only available in the context of a simulation, we can use for reference a state $S_{i'}$, which is the state existing in a parallel chain in which every transition was using an alternative fair transition.

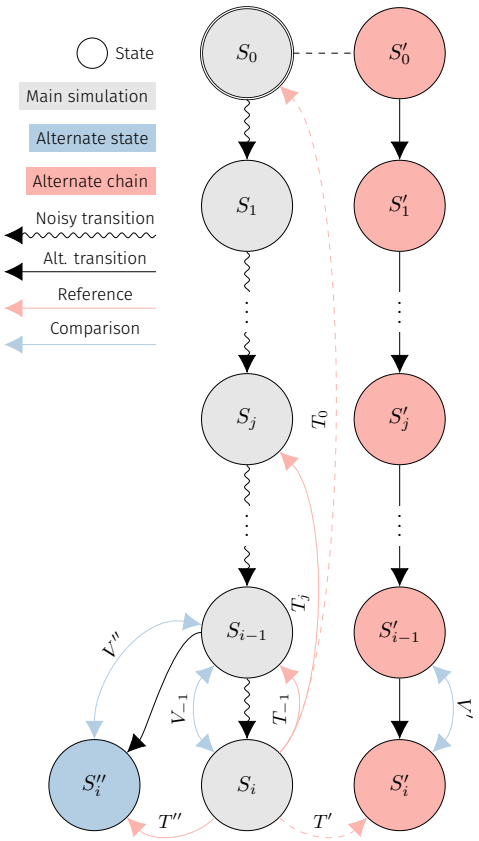


Figure 5.15.: Scenarios

5.6.4. Red flag notion of fairness

The idea of "automated 'early warning system' for discrimination" is proposed in [Ben23], and we share this objective with this model. This requires the definition of a "warning", therefore needing an indicator to be selected. However, selecting one fairness metric above the others would go against this thesis. As an agnostic fairness metric–if any–we advocate for a distance between distributions, such as the earth mover's distance for instance. This does not provide a fairness notion by itself–outside of the extreme case of requiring everyone to receive the same decision, but can be used as a "red flag" indicating a *potential* issue to be investigated. Indeed, it would be "raised" when disparity arises between the distributions *of both attributes and decisions* across groups, alerting experts of the field that a deeper inspection is required to ensure that the mechanism under scrutiny is fair. Rather than taking a state as reference as discussed above, we would then need to compare the evolution between two states of time, assessing the rate by which the distance increases (or, ideally, decreases) and comparing that rate to another one. This is represented with blue arrows on Figure 5.15, in which the distribution changes V_{-1} (between states S_i and S_{i-1}) can be compared to the change V'' or V'. This approach would allow to define a "(un)fairness rate", which corresponds to a direction (i.e., increase or decrease of distance) and velocity of (un)fairness. Rather than adding

to the existing fairness metrics, this could serve as a way to detect if an investigation by proper experts is *potentially* needed.

6. Conclusion

In this thesis, we have attempted to highlight the intrinsic difficulties of applying technical solutions to social issues, and how these solutions appear unsuited to the actual concern. More precisely, first we have looked at the difficulties encountered when state-of-the-art privacy-preserving mechanisms need to be applied under strong–if not immutable–utility and transparency constraints. This clash is arising with the will to develop technological tools benefiting from data that is, at the same time, subject to social requirements, as is the case with legal proceedings.

Afterwards, we shifted towards the evaluation of privacy empirical guarantees, with the objective of better understanding how formal frameworks would impact real-life data and processes. This was done by creating a framework to help with the instantiation of attack challenges, and by organizing a competition on the issue of membership inference attacks of privacy-preserving synthetic data. Our hope is to organize further editions, and potentially to tackle further questions such as the aforementioned trade-off between privacy and utility of natural language sanitization, or the study of the fairness impact of privacy mechanisms.

As a last contribution, we finally turned to fairness issues with the objective of proposing a framework to observe long-term dynamics of technical fairness notions. To this end, we reproduced and adapted a model originating from the community of education policy, with the aim of providing a technical tool grounded in reality. This last point should be thought as both the ending note of this thesis and the broad direction that we identify for the future of this work: identifying not how to use social concepts in technical contexts, but rather identifying when and how technical tools can be used to help with those social issues. Therefore, keeping social issues in the domain of social sciences, and avoiding the technicization of social concepts as if the "technical" was outside the sociotechnical system.

Bibliography

[AC08] Ahmed Abbasi and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace". In: *ACM Trans. Inf. Syst.* 26.2 (2008), 7:1–7:29. DOI: 10.1145/1344411.1344413 (cit. on p. 35).

[AGH70] James R. Abernathy, Bernard G. Greenberg, and Daniel G. Horvitz. "Estimates of induced abortion in urban North Carolina". In: *Demography* 7.1 (1970), pp. 19–29. DOI: 10.2307/2060019 (cit. on p. 23).

[Abo18] John M. Abowd. "The U.S. Census Bureau Adopts Differential Privacy". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. Ed. by Yike Guo and Faisal Farooq. ACM, 2018, p. 2867. DOI: 10.1145/3219819.3226070 (cit. on pp. 22, 52).

[ACM18] ACM, ed. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*. 2018 (cit. on pp. 4, 13).

[Aïv+21] Ulrich Aïvodji et al. "Characterizing the risk of fairwashing". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato et al. 2021, pp. 14822–14834 (cit. on p. 28).

[Akt+20] Ahmet Aktay et al. "Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.0)". In: *CoRR* (2020). arXiv: 2004.04145 (cit. on p. 22).

[ANY17] Benjamin Alarie, Anthony Niblett, and Albert H Yoon. "How Artificial Intelligence Will Affect the Practice of Law". In: *University of Toronto Law Journal* 68 (supplement 1 2017), pp. 106–124. DOI: 10.2139/ssrn.3066816 (cit. on p. 34).

[Ale+16] Nikolaos Aletras et al. "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective". In: *PeerJ Comput. Sci.* 2 (2016), e93. DOI: 10.7717/peerj-cs.93 (cit. on pp. 33, 34).

[ABG20a] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "Online publication of court records: circumventing the privacy-transparency trade-off". In: *CoRR* (2020). Presented at ICML 2020's Law and Machine Learning Workshop. arXiv: 2007.01688 (cit. on pp. 9, 17, 30).

[ABG20b] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "Publication of Court Records: Circumventing the Privacy-Transparency Trade-Off". In: *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers*. Ed. by Víctor Rodríguez-Doncel et al. Vol. 13048. Lecture Notes in Computer Science. Springer, 2020, pp. 298–312. DOI: 10.1007/978-3-030-89811-3_21. HAL: hal-03225201 (cit. on pp. 9, 17, 22, 30).

[ABG23a] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "[~Re] Simulating Socioeconomic-Based Affirmative Action". In: *ReScience C* 9.1 (2023). Ed. by Olivia Guest. DOI: 10.5281/zenodo.10255346. HAL: hal-04328511 (cit. on pp. 10, 19, 58).

[ABG23b] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "1st Edition of the SNAKE Challenge. SNAKE #1: SaNitization Algorithm under attacK ϵ". collocated with the 13ième Atelier sur la Protection de la Vie Privée (APVP'23). 2023 (cit. on pp. 8, 9, 18, 44).

[ABG23c] Tristan Allard, Louis Béziaud, and Sébastien Gambs. "SNAKE Challenge: Sanitization Algorithms under Attack". In: *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management*. Birmingham, United Kingdom: Association for Computing Machinery, Oct. 21, 2023. DOI: 10.1145/3583780.3614754. HAL: hal-04228115 (cit. on pp. 8, 9, 18, 44).

[AGB21] Tristan Allard, Sébastien Gambs, and Louis Béziaud. "La Confidentialité Différentielle, Garante de l'anonymat". In: *Pour la Science*. Hors-Série 112 (July 8, 2021).

[All+20] Tristan Allard et al. "Ouvrir La Boîte Noire Des Algorithmes de Personnalisation". In: *Le Profilage En Ligne : Entre Libéralisme et Régulation*. Ed. by Alexandra Bensamoun, Maryline Boizard, and Sandrine Turgis. Libre Droit. Mare & Martin, Oct. 15, 2020, pp. 211–231. ISBN: 978-2-84934-466-8.

[Alv+18] Mário S. Alvim et al. "Metric-based local differential privacy for statistical applications". In: *CoRR* (2018). arXiv: 1805.01456 (cit. on p. 42).

[AC11] Balamurugan Anandan and Chris Clifton. "Significance of Term Relationships on Anonymization". In: *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2011, Campus Scientifique de la Doua, Lyon, France, August 22-27, 2011*. Ed. by Jomi Fred Hübner, Jean-Marc Petit, and Einoshin Suzuki. IEEE Computer Society, 2011, pp. 253–256. DOI: 10.1109/WI-IAT.2011.240 (cit. on p. 38).

[Ara13] Ayala Arad. "Past decisions do affect future choices: An experimental demonstration". In: *Organizational Behavior and Human Decision Processes* 121.2 (July 2013), pp. 267–277. DOI: 10.1016/j.obhdp.2013.01.006 (cit. on p. 14).

[Arr06] Michael Arrington. "AOL Proudly Releases Massive Amounts of Private Data". In: *TechCrunch* (2006) (cit. on p. 35).

[AB09] Kevin D. Ashley and Stefanie Brüninghaus. "Automatically classifying case texts and predicting outcomes". In: *Artif. Intell. Law* 17.2 (2009), pp. 125–165. DOI: 10.1007/s10506-009-9077-9 (cit. on p. 34).

[AW13] Kevin D. Ashley and Vern R. Walker. "Toward constructing evidence-based legal arguments using legal decision documents and machine learning". In: *International Conference on Artificial Intelligence and Law, ICAIL '13, Rome, Italy, June 10-14, 2013*. Ed. by Enrico Francesconi and Bart Verheij. ACM, 2013, pp. 176–180. DOI: 10.1145/2514601.2514622 (cit. on p. 34).

[Ata+00] Mikhail J. Atallah et al. "Natural language processing for information assurance and security: an overview and implementations". In: *Proceedings of the 2000 Workshop on New Security Paradigms, Ballycotton, Co. Cork, Ireland, September 18-21, 2000*. Ed. by Mary Ellen Zurko and Steven J. Greenwald. ACM, 2000, pp. 51–65. DOI: 10.1145/366173.366190 (cit. on p. 40).

[Ate+15] Giuseppe Ateniese et al. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers". In: *Int. J. Secur. Networks* 10.3 (2015), pp. 137–150. DOI: 10.1504/IJSN.2015.071829 (cit. on p. 24).

[Ave11] Terje Aven. "Selective critique of risk assessments with recommendations for improving methodology and practise". In: *Reliab. Eng. Syst. Saf.* 96.5 (2011), pp. 509–514. DOI: 10.1016/j.ress.2010.12.021 (cit. on pp. 4, 13).

[BB16] Jane Bailey and Jacquelyn Burkell. "Revisiting the Open Court Principle in an Era of Online Publication: Questioning Presumptive Public Access to Parties' and Witnesses' Personal Information". In: *Ottawa L. Rev.* 48 (2016), p. 143 (cit. on p. 31).

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019 (cit. on p. 29).

[Bau+21] Greta R. Bauer et al. "Intersectionality in Quantitative Research: A Systematic Review of Its Emergence and Applications of Theory and Methods". In: *SSM - Population Health* 14 (2021). DOI: 10.1016/j.ssmph.2021.100798 (cit. on pp. 5, 17).

[Bel+19] Rachel K. E. Bellamy et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM J. Res. Dev.* 63.4/5 (2019), 4:1–4:15. DOI: 10.1147/JRD.2019.2942287 (cit. on pp. 27, 69).

[Ben23] Bilel Benbouzid. "Fairness in machine learning from the perspective of sociology of statistics: How machine learning is becoming scientific by turning its back on metrological realism". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*. ACM, 2023, pp. 35–43. DOI: 10.1145/3593013.3593974 (cit. on pp. 21, 28, 70).

[BB43] Jeremy Bentham and John Bowring. *The Works of Jeremy Bentham*. Vol. 4. W. Tait, 1843 (cit. on p. 30).

[BP08] Alain Bihr and Roland Pfefferkorn. "III. Le Cumul Des Inégalités". In: *Reperes* (2008), pp. 55–77 (cit. on p. 15).

[Bir+22] Abeba Birhane et al. "The Forgotten Margins of AI Ethics". In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 2022, pp. 948–958. DOI: 10.1145/3531146.3533157 (cit. on pp. 4, 13).

[BP64] Pierre Bourdieu and Jean-Claude Passeron. *Les Héritiers. Les Étudiants et La Culture*. Les éditions de Minuit, 1964 (cit. on p. 14).

[BP70] Pierre Bourdieu and Jean-Claude Passeron. *La Reproduction. Éléments Pour Une Théorie Du Système d'enseignement*. Les Éditions de Minuit, 1970 (cit. on p. 14).

[Bou+20] Antoine Boutet et al. "DARC : Data Anonymization and Re-Identification Challenge". In: *RESSI 2020 - Rendez-Vous de La Recherche et de l'Enseignement de La Sécurité Des Systèmes d'Information*. 2020. HAL: hal-02512677 (cit. on p. 44).

[BD18] Tim Brennan and William Dieterich. "Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)". In: *Handbook of Recidivism Risk/Needs Assessment Tools*. John Wiley & Sons, Ltd, 2018. Chap. 3, pp. 49–75. ISBN: 9781119184256. DOI: 10.1002/9781119184256.ch3 (cit. on pp. 4, 13, 14).

[Bro+16] Greg Brockman et al. "OpenAI Gym". In: *CoRR* (2016). arXiv: 1606.01540 (cit. on p. 27).

[BA97] Stefanie Brüninghaus and Kevin D. Ashley. "Using Machine Learning for Assigning Indices to Textual Cases". In: *Case-Based Reasoning Research and Development, Second International Conference, ICCBR-97, Providence, Rhode Island, USA, July 25-27, 1997, Proceedings*. Ed. by David B. Leake and Enric Plaza. Vol. 1266. Lecture Notes in Computer Science. Springer, 1997, pp. 303–314. DOI: 10.1007/3-540-63233-6_501 (cit. on p. 34).

[Cal14] Krishnadev Calamur. "In a First for Britain, a Secret Trial for Terrorism Suspects". In: *NPR* (June 5, 2014) (cit. on p. 31).

[Can79] CanLII, ed. *Katopodis v. Katopodis*. 1979 (cit. on p. 38).

[Can15] CanLII, ed. *Droit de La Famille – 15334, 2015 QCCS 762*. 2015 (cit. on p. 37).

[CW22] Alycia N. Carey and Xintao Wu. "The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences". In: *Frontiers Big Data* 5 (2022), p. 892837. DOI: 10.3389/fdata.2022.892837 (cit. on pp. 26, 27).

[CS21] Hongyan Chang and Reza Shokri. "On the Privacy Risks of Algorithmic Fairness". In: *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*. IEEE, 2021, pp. 292–303. DOI: 10.1109/EuroSP51992.2021.00028 (cit. on p. 29).

[Che+09] Bee-Chung Chen et al. "Privacy-Preserving Data Publishing". In: *Found. Trends Databases* 2.1-2 (2009), pp. 1–167. DOI: 10.1561/1900000008 (cit. on p. 35).

[Cho17] Alexandra Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2 (2017), pp. 153–163. DOI: 10.1089/big.2016.0047 (cit. on p. 29).

[CR20] Alexandra Chouldechova and Aaron Roth. "A snapshot of the frontiers of fairness in machine learning". In: *Commun. ACM* 63.5 (2020), pp. 82–89. DOI: 10.1145/3376898 (cit. on p. 28).

[CT13] Chris Clifton and Tamir Tassa. "On Syntactic Anonymity and Differential Privacy". In: *Trans. Data Priv.* 6.2 (2013), pp. 161–183 (cit. on p. 15).

[Cod23] CodeX, the Stanford Center for Legal Informatics, ed. *Legaltechlist*. 2023 (cit. on p. 33).

[Con+11] Amanda Conley et al. "Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry". In: *Md. L. Rev.* 71 (2011), p. 772 (cit. on pp. 31, 33, 36).

[Cre89] Kimberlé Williams Crenshaw. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: *Feminist legal theories*. 1989 (cit. on pp. 4, 17).

[Cum+19] Rachel Cummings et al. "On the Compatibility of Privacy and Fairness". In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 09-12, 2019*. Ed. by George Angelos Papadopoulos et al. ACM, 2019, pp. 309–315. DOI: 10.1145/3314183.3323847 (cit. on p. 28).

[Cus+19] Tonya Custis et al. "Westlaw Edge AI Features Demo: KeyCite Overruling Risk, Litigation Analytics, and WestSearch Plus". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*. ACM, 2019, pp. 256–257. DOI: 10.1145/3322640.3326739 (cit. on p. 34).

[DAm+20] Alexander D'Amour et al. "Fairness is not static: deeper understanding of long term fairness via simulation studies". In: *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt et al. ACM, 2020, pp. 525–534. DOI: 10.1145/3351095.3372878 (cit. on p. 27).

[Dal19] Robert Dale. "Law and Word Order: NLP in Legal Tech". In: *Nat. Lang. Eng.* 25.1 (2019), pp. 211–217. DOI: 10.1017/S1351324918000475 (cit. on p. 33).

[Dal77] Tore Dalenius. "Towards a Methodology for Statistical Disclosure Control". In: *Journal of Official Statistics* (1977) (cit. on p. 21).

[Dal86] Tore Dalenius. "Finding a Needle in a Haystack or Identifying Anonymous Census Records". In: *Journal of official statistics* 2.3 (1986), p. 329. DOI: 10.1111/nyas.13660 (cit. on p. 20).

[Das18] Jeffrey Dastin. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women". In: *Reuters* (Oct. 2018) (cit. on pp. 4, 13).

[DP20] Damien Desfontaines and Balázs Pejó. "SoK: Differential privacies". In: *Proc. Priv. Enhancing Technol.* 2020.2 (2020), pp. 288–313. DOI: 10.2478/popets-2020-0028 (cit. on pp. 5, 16, 21, 23).

[DM23] Fernando Diaz and Michael Madaio. "Scaling Laws Do Not Scale". In: *CoRR* (2023). DOI: 10.48550/arXiv.2307.03201. arXiv: 2307.03201 (cit. on p. 22).

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. "Collecting Telemetry Data Privately". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 3571–3580 (cit. on p. 22).

[Dou+05] M.M. Douglass et al. "De-identification algorithm for free-text nursing notes". In: *Computers in Cardiology, 2005.* 2005, pp. 331–334. DOI: 10.1109/CIC.2005.1588104 (cit. on p. 40).

[Dwo06] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*. Ed. by Michele Bugliesi et al. Vol. 4052. Lecture Notes in Computer Science. Springer, 2006, pp. 1–12. DOI: 10.1007/11787006_1 (cit. on pp. 5, 15, 22).

[DN92] Cynthia Dwork and Moni Naor. "Pricing via Processing or Combatting Junk Mail". In: *Advances in Cryptology - CRYPTO '92, 12th Annual International Cryptology Conference, Santa Barbara, California, USA, August 16-20, 1992, Proceedings*. Ed. by Ernest F. Brickell. Vol. 740. Lecture Notes in Computer Science. Springer, 1992, pp. 139–147. DOI: 10.1007/3-540-48071-4_10 (cit. on p. 41).

[DR14] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3-4 (2014), pp. 211–407. DOI: 10.1561/0400000042 (cit. on pp. 8, 22, 23, 42, 45).

[Dwo+12] Cynthia Dwork et al. "Fairness through awareness". In: *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. Ed. by Shafi Goldwasser. ACM, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255 (cit. on pp. 6, 16, 26).

[Eco23] Economic Policy Institute. *Current Population Survey Extracts*. Version 1.0.39. 2023 (cit. on p. 48).

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*. Ed. by Gail-Joon Ahn, Moti Yung, and Ninghui Li. ACM, 2014, pp. 1054–1067. DOI: 10.1145/2660267.2660348 (cit. on p. 22).

[Eur21] European Commission, Directorate-General for Communications Networks, Content and Technology. *Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. 2021 (cit. on pp. 4, 13).

[Eur16] European Parliament. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)". In: *OJ* (May 4, 2016), pp. 1–88 (cit. on pp. 4, 13, 15).

[FES22] Jake Fawkes, Robin J. Evans, and Dino Sejdinovic. "Selection, Ignorability and Challenges With Causal Fairness". In: *1st Conference on Causal Learning and Reasoning, CLeaR 2022, Sequoia Conference Center, Eureka, CA, USA, 11-13 April, 2022*. Ed. by Bernhard Schölkopf, Caroline Uhler, and Kun Zhang. Vol. 177. Proceedings of Machine Learning Research. PMLR, 2022, pp. 275–289 (cit. on p. 28).

[Fed23] Federation of Law Societies of Canada, ed. *Canadian Legal Information Institute*. 2023 (cit. on p. 33).

[Fel+15] Michael Feldman et al. "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. Ed. by Longbing Cao et al. ACM, 2015, pp. 259–268. DOI: 10.1145/2783258.2783311 (cit. on pp. 26, 27).

[FDM19] Natasha Fernandes, Mark Dras, and Annabelle McIver. "Generalised Differential Privacy for Text Document Processing". In: *Principles of Security and Trust - 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings*. Ed. by Flemming Nielson and David Sands. Vol. 11426. Lecture Notes in Computer Science. Springer, 2019, pp. 123–148. DOI: 10.1007/978-3-030-17138-4_6 (cit. on pp. 40, 42).

[Fin93] Craig A. Finseth. "The Uniqueness of Unique Identifiers". In: *RFC* 1439 (1993), pp. 1–11. DOI: 10.17487/RFC1439 (cit. on p. 37).

[Fio+22] Ferdinando Fioretto et al. "Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. Ed. by Luc De Raedt. ijcai.org, 2022, pp. 5470–5477. DOI: 10.24963/ijcai.2022/766 (cit. on pp. 28, 29).

[Fle21] Will Fleisher. "What's Fair about Individual Fairness?" In: *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. Ed. by Marion Fourcade et al. ACM, 2021, pp. 480–490. DOI: 10.1145/3461702.3462621 (cit. on p. 28).

[Fle17] Caroline Fleuriot. "Avec l'accès Gratuit à Toute La Jurisprudence, Des Magistrats Réclament l'anonymat". In: *Dalloz Actualité* (Feb. 2017) (cit. on p. 31).

[Fou+20] James R. Foulds et al. "An Intersectional Definition of Fairness". In: *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*. IEEE, 2020, pp. 1918–1921. DOI: 10.1109/ICDE48307.2020.00203 (cit. on pp. 5, 17, 27).

[Fou+19] Amaury Fouret et al. *Open Justice*. Tech. rep. Cour de cassation, 2019 (cit. on p. 31).

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*. Ed. by Indrajit Ray, Ninghui Li, and Christopher Kruegel. ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677 (cit. on p. 24).

[Fre23] Free Law Project, ed. *CourtListener*. 2023 (cit. on p. 33).

[FSV16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im)possibility of fairness". In: *CoRR* (2016). arXiv: 1609.07236 (cit. on pp. 6, 16, 29).

[FPZ19] Leïla Frouillou, Clément Pin, and Agnès van Zanten. "Le Rôle Des Instruments Dans La Sélection Des Bacheliers Dans l'enseignement Supérieur. La Nouvelle Gouvernance Des Affectations Par Les Algorithmes". In: *Sociologie* 10.2 (2019), p. 209. DOI: 10.3917/socio.102.0209 (cit. on pp. 4, 13).

[GB20] David García-Soriano and Francesco Bonchi. "Fair-by-design matching". In: *Data Min. Knowl. Discov.* 34.5 (2020), pp. 1291–1335. DOI: 10.1007/s10618-020-00675-y (cit. on pp. 26, 27).

[Goo+20] Ian J. Goodfellow et al. "Generative adversarial networks". In: *Commun. ACM* 63.11 (2020), pp. 139–144. DOI: 10.1145/3422622 (cit. on p. 49).

[GG75] Michael S Goodstadt and Valerie Gruson. "The Randomized Response Technique: A Test on Drug Use". In: *Journal of the American Statistical Association* 70.352 (1975), pp. 814–818 (cit. on p. 23).

[Gre+69] Bernard G. Greenberg et al. "The Unrelated Question Randomized Response Model: Theoretical Framework". In: *Journal of the American Statistical Association* 64 (1969), pp. 520–539 (cit. on p. 23).

[GG07] Bernard Guerin and Pauline Guerin. "Lessons Learned from Participatory Discrimination Research: Long-term Observations and Local Interventions". PhD thesis. College of Community Psychologists of the Australian Psychological Society, 2007 (cit. on p. 52).

[GK21] Swati Gupta and Vijay Kamble. "Individual Fairness in Hindsight". In: *J. Mach. Learn. Res.* 22 (2021), 144:1–144:35 (cit. on pp. 26, 28).

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Daniel D. Lee et al. 2016, pp. 3315–3323 (cit. on p. 26).

[HS13] Woodrow Hartzog and Frederic Stutzman. "The Case for Online Obscurity". In: *Calif. L. Rev.* 101 (2013), p. 1 (cit. on p. 41).

[Has+19] Fadi Hassan et al. "Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings". In: *18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 13th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2019, Rotorua, New Zealand, August 5-8, 2019*. IEEE, 2019, pp. 358–365. DOI: 10.1109/TrustCom/BigDataSE.2019.00055 (cit. on p. 41).

[HZL19] Zecheng He, Tianwei Zhang, and Ruby B. Lee. "Model inversion attacks against collaborative inference". In: *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019*. Ed. by David Balenson. ACM, 2019, pp. 148–162. DOI: 10.1145/3359789.3359824 (cit. on p. 24).

[Hel+22] Jonas Helgertz et al. “A New Strategy for Linking US Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel”. In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55.1 (2022), pp. 12–29. DOI: 10.1080/01615440.2021.1985027 (cit. on p. 52).

[HHB19] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. “Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models”. In: *Proc. Priv. Enhancing Technol.* 2019.4 (2019), pp. 232–249. DOI: 10.2478/popets-2019-0067 (cit. on p. 47).

[Hof19] Anna Lauren Hoffmann. “Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse”. In: *Information, Communication & Society* 22 (2019), pp. 900–915. DOI: 10.1080/1369118X.2019.1573912 (cit. on pp. 4, 13, 29).

[Hou+22] Florimond Houssiau et al. “TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data”. In: *CoRR* (2022). DOI: 10.48550/arXiv.2211.06550. arXiv: 2211.06550 (cit. on p. 51).

[HH17] Stephan Hoyer and Joe Hamman. “Xarray: ND Labeled Arrays and Datasets in Python”. In: *Journal of Open Research Software* 5.1 (2017). DOI: DOI: 10.5334/jors.148 (cit. on p. 58).

[Hsu+14] Justin Hsu et al. “Differential Privacy: An Economic Method for Choosing Epsilon”. In: *IEEE 27th Computer Security Foundations Symposium, CSF 2014, Vienna, Austria, 19-22 July, 2014*. IEEE Computer Society, 2014, pp. 398–410. DOI: 10.1109/CSF.2014.35 (cit. on pp. 5, 16, 18).

[Hu+21] Aoting Hu et al. “TableGAN-MCA: Evaluating Membership Collisions of GAN-Synthesized Tabular Data Releasing”. In: *CCS ’21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*. Ed. by Yongdae Kim et al. ACM, 2021, pp. 2096–2112. DOI: 10.1145/3460120.3485251 (cit. on p. 47).

[Hu+22] Hongsheng Hu et al. “Membership Inference Attacks on Machine Learning: A Survey”. In: *ACM Comput. Surv.* 54.11s (2022), 235:1–235:37. DOI: 10.1145/3523273 (cit. on pp. 8, 24, 45).

[HIV19] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. “The Disparate Effects of Strategic Manipulation”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 259–268. DOI: 10.1145/3287560.3287597 (cit. on p. 27).

[Hye+22] Jihyeon Hyeong et al. “An Empirical Study on the Membership Inference Attack against Tabular Data Synthesis Models”. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. Ed. by Mohammad Al Hasan and

Li Xiong. ACM, 2022, pp. 4064–4068. DOI: 10.1145/3511808.3557546 (cit. on p. 47).

[Jac02] Joseph Jaconelli. *Open Justice: A Critique of the Public Trial*. Oxford University Press on Demand, 2002 (cit. on p. 31).

[Jac17] Jean-Baptiste Jacquin. "Terrorisme : La Peur Des Magistrats". In: *Le Monde* (Jan. 2017) (cit. on p. 31).

[JUO20] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. "Auditing Differentially Private Machine Learning: How Private is Private SGD?" In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 22205–22216 (cit. on p. 24).

[Jia+09] Wei Jiang et al. "t-Plausibility: Semantic Preserving Text Sanitization". In: *Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, CSE 2009, Vancouver, BC, Canada, August 29-31, 2009*. IEEE Computer Society, 2009, pp. 68–75. DOI: 10.1109/CSE.2009.353 (cit. on p. 37).

[Joa98] Thorsten Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In: *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*. Ed. by Claire Nedellec and Céline Rouveirol. Vol. 1398. Lecture Notes in Computer Science. Springer, 1998, pp. 137–142. DOI: 10.1007/BFb0026683 (cit. on p. 34).

[JYS19] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019 (cit. on pp. 42, 49).

[Jor+20] James Jordon et al. "Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-identification". In: *NeurIPS 2020 Competition and Demonstration Track, 6-12 December 2020, Virtual Event / Vancouver, BC, Canada*. Ed. by Hugo Jair Escalante and Katja Hofmann. Vol. 133. Proceedings of Machine Learning Research. PMLR, 2020, pp. 206–215 (cit. on pp. 8, 18, 44).

[Jou+17] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics, 2017, pp. 427–431. DOI: 10.18653/v1/e17-2068 (cit. on p. 34).

[Jud03] Judges Technology Advisory Committee. *Open Courts, Electronic Access to Court Records, and Privacy: Discussion Paper*. Tech. rep. Canadian Judicial Council, 2003 (cit. on p. 37).

[KBK14] Hiroshi Kajino, Yukino Baba, and Hisashi Kashima. "Instance-Privacy Preserving Crowdsourcing". In: *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*. Ed. by Jeffrey P. Bigham and David C. Parkes. AAAI, 2014, pp. 96–103. DOI: 10.1609/hcomp.v2i1.13146 (cit. on p. 43).

[KC11] Faisal Kamiran and Toon Calders. "Data preprocessing techniques for classification without discrimination". In: *Knowl. Inf. Syst.* 33.1 (2011), pp. 1–33. DOI: 10.1007/s10115-011-0463-8 (cit. on p. 27).

[Kan+22] Jian Kang et al. "InfoFair: Information-Theoretic Intersectional Fairness". In: *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*. Ed. by Shusaku Tsumoto et al. IEEE, 2022, pp. 1455–1464. DOI: 10.1109/BigData55660.2022.10020588 (cit. on pp. 5, 17).

[Kas+11] Shiva Prasad Kasiviswanathan et al. "What Can We Learn Privately?" In: *SIAM J. Comput.* 40.3 (2011), pp. 793–826. DOI: 10.1137/090756090 (cit. on p. 23).

[KBB17] Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. "A general approach for predicting the behavior of the Supreme Court of the United States". In: *PLoS One* 12.4 (2017). Ed. by Luís A. Nunes Amaral, e0174698. DOI: 10.1371/journal.pone.0174698 (cit. on pp. 33, 40).

[Kea+19] Michael J. Kearns et al. "An Empirical Study of Rich Subgroup Fairness for Machine Learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 100–109. DOI: 10.1145/3287560.3287592 (cit. on p. 27).

[KRG19] Mi-Young Kim, Juliano Rabelo, and Randy Goebel. "Statute Law Information Retrieval and Entailment". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*. ACM, 2019, pp. 283–289. DOI: 10.1145/3322640.3326742 (cit. on p. 34).

[Kim14] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, 2014, pp. 1746–1751. DOI: 10.3115/v1/d14-1181 (cit. on p. 34).

[Kor65] Fred Kort. "Quantitative Analysis of Fact-Patterns in Cases and Their Impact on Judicial Decisions". In: *Harv. L. Rev.* 79 (1965), p. 1595 (cit. on p. 34).

[KR18] Johannes Köster and Sven Rahmann. “Snakemake - a scalable bioinformatics workflow engine”. In: *Bioinform.* 34.20 (2018), p. 3600. DOI: 10.1093/bioinformatics/bty350 (cit. on p. 45).

[KK16] Oliver Kramer and Oliver Kramer. “Scikit-Learn”. In: *Machine learning for evolution strategies* (2016), pp. 45–53 (cit. on p. 58).

[Kri+20] Kalpesh Krishna et al. “Thieves on Sesame Street! Model Extraction of BERT-based APIs”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020 (cit. on p. 24).

[Kul+20] Bogdan Kulynych et al. “POTs: protective optimization technologies”. In: *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt et al. ACM, 2020, pp. 177–188. DOI: 10.1145/3351095.3372853 (cit. on pp. 4, 13).

[Kus+17] Matt J. Kusner et al. “Counterfactual Fairness”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 4066–4076 (cit. on p. 27).

[Lai+15] Siwei Lai et al. “Recurrent Convolutional Neural Networks for Text Classification”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. Ed. by Blai Bonet and Sven Koenig. AAAI Press, 2015, pp. 2267–2273. DOI: 10.1609/aaai.v29i1.9513 (cit. on p. 34).

[LC11] Jaewoo Lee and Chris Clifton. “How Much Is Enough? Choosing ϵ for Differential Privacy”. In: *Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings*. Ed. by Xuejia Lai, Jianying Zhou, and Hui Li. Vol. 7001. Lecture Notes in Computer Science. Springer, 2011, pp. 325–340. DOI: 10.1007/978-3-642-24861-0_22 (cit. on pp. 5, 16).

[LDR06] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. “Mondrian Multidimensional K-Anonymity”. In: *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*. Ed. by Ling Liu et al. IEEE Computer Society, 2006, p. 25. DOI: 10.1109/ICDE.2006.101 (cit. on p. 20).

[Leg09] Legalis, ed. *Cour d'appel de Paris 11ème Chambre, Section B Arrêt Du 14 Février 2008*. 2009 (cit. on p. 38).

[Lég13] Légifrance, ed. *Cour de Cassation, Civile, Chambre Civile 1, 10 Avril 2013, 12-14.525, Publié Au Bulletin*. 2013 (cit. on p. 39).

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”. In: *2007 IEEE 23rd International Conference on Data Engineering*. 2007, pp. 106–115. DOI: 10.1109/ICDE.2007.367856 (cit. on pp. 5, 15, 20).

[Liu+19] Lydia T. Liu et al. “Delayed Impact of Fair Machine Learning”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 6196–6200. DOI: 10.24963/ijcai.2019/862 (cit. on p. 27).

[LQH16] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. “Recurrent Neural Network for Text Classification with Multi-Task Learning”. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. Ed. by Subbarao Kambhampati. IJCAI/AAAI Press, 2016, pp. 2873–2879 (cit. on p. 34).

[Mac+07] Ashwin Machanavajjhala et al. “L-diversity: Privacy beyond k-anonymity”. In: *ACM Trans. Knowl. Discov. Data* 1.1 (2007), p. 3. DOI: 10.1145/1217299.1217302 (cit. on pp. 5, 15, 20).

[ML21] Abdul Majeed and Sungchang Lee. “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey”. In: *IEEE Access* 9 (2021), pp. 8512–8545. DOI: 10.1109/ACCESS.2020.3045700 (cit. on p. 35).

[Man+17] Arpan Mandal et al. “Measuring Similarity among Legal Court Case Documents”. In: *Proceedings of the 10th Annual ACM India Compute Conference, Compute 2017, Bhopal, India, November 16-18, 2017*. Ed. by Partha Pratim Chakraborty et al. ACM, 2017, pp. 1–9. DOI: 10.1145/3140107.3140119 (cit. on p. 33).

[Mar+19] Max Raphael Sobroza Marques et al. “Machine learning for explaining and ranking the most influential matters of law”. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*. ACM, 2019, pp. 239–243. DOI: 10.1145/3322640.3326734 (cit. on p. 34).

[Mar+13] Mónica Marrero et al. “Named Entity Recognition: Fallacies, challenges and opportunities”. In: *Comput. Stand. Interfaces* 35.5 (2013), pp. 482–489. DOI: 10.1016/j.csi.2012.09.004 (cit. on p. 31).

[Mar08] Peter W Martin. “Online Access to Court Records-from Documents to Data, Particulars to Patterns”. In: *Vill. L. Rev.* 53 (2008), p. 855 (cit. on p. 33).

[McC11] Tom McClean. “Not with a Bang but a Whimper: The Politics of Accountability and Open Data in the UK”. In: *APSA 2011 Annual Meeting Paper*. 2011 (cit. on p. 30).

[McD10] Patrice McDermott. “Building Open Government”. In: *Government Information Quarterly* 27.4 (2010), pp. 401–413. DOI: 10.1016/j.giq.2010.07.002 (cit. on p. 30).

[MMS21] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. "Winning the NIST Contest: A scalable and general approach to differentially private synthetic data". In: *J. Priv. Confidentiality* 11.3 (2021). DOI: 10.29012/jpc.778 (cit. on pp. 42, 47, 49).

[MSM19] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. "Graphical-model based estimation and inference for differential privacy". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 4435–4444 (cit. on p. 49).

[Meh+22] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6 (2022), 115:1–115:35. DOI: 10.1145/3457607 (cit. on pp. 6, 16, 26, 27).

[Mik+13] Tomás Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges et al. 2013, pp. 3111–3119 (cit. on p. 34).

[MS18] Akshay Minocha and Navjyoti Singh. "Legal Document Similarity Using Triples Extracted from Unstructured Text". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Georg Rehm, Víctor Rodríguez-Doncel, and Julián Moreno-Schneider. Paris, France: European Language Resources Association (ELRA), May 2018. ISBN: 979-10-95546-18-4 (cit. on p. 33).

[Mit+86] Tom M Mitchell et al. "Judge: A Case-Based Reasoning System". In: *Machine Learning: A Guide to Current Research* (1986), pp. 1–4 (cit. on p. 14).

[MSC19] Ivan Mokanov, Daniel Shane, and Benjamin Cerat. "Facts2Law: using deep learning to provide a legal qualification to a set of facts". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*. ACM, 2019, pp. 268–269. DOI: 10.1145/3322640.3326694 (cit. on pp. 34, 35).

[Mol23] Callum Mole. *Reprosyn*. 2023 (cit. on p. 49).

[ML22] Mathieu Molina and Patrick Loiseau. "Bounding and Approximating Intersectional Fairness through Marginal Fairness". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 16796–16807 (cit. on pp. 5, 17).

[MFL10] Mehdi Yousfi Monod, Atefeh Farzindar, and Guy Lapalme. "Supervised Machine Learning for Summarizing Legal Documents". In: *Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence, Canadian, AI 2010, Ottawa, Canada, May 31 - June 2, 2010. Proceedings.* Ed. by Atefeh Farzindar and Vlado Keselj. Vol. 6085. Lecture Notes in Computer Science. Springer, 2010, pp. 51–62. DOI: 10.1007/978-3-642-13059-5_8 (cit. on p. 34).

[Mor+19] Giulio Morina et al. "Auditing and Achieving Intersectional Fairness in Classification Problems". In: *CoRR* (2019). arXiv: 1911.01468 (cit. on pp. 5, 17).

[Moz+05] Martin Mozina et al. "Argument Based Machine Learning Applied to Law". In: *Artif. Intell. Law* 13.1 (2005), pp. 53–73. DOI: 10.1007/s10506-006-9002-4 (cit. on p. 34).

[Mur+23] Takao Murakami et al. "Designing a Location Trace Anonymization Contest". In: *Proc. Priv. Enhancing Technol.* 2023.1 (2023), pp. 225–243. DOI: 10.56553/popets-2023-0014 (cit. on p. 44).

[NM08] Ramesh Nallapati and Christopher D. Manning. "Legal Docket Classification: Where Machine Learning Stumbles". In: *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL.* ACL, 2008, pp. 438–446 (cit. on p. 34).

[Nar18] Arvind Narayanan. "Tutorial: 21 Fairness Definitions and Their Politics". In: *Proc. Conf. Fairness Accountability Transp., New York, USA.* Vol. 1170. 2018 (cit. on pp. 6, 16).

[Nas90] Roy Nash. "Bourdieu on Education and Social and Cultural Reproduction". In: *British journal of sociology of education* 11.4 (1990), pp. 431–447. DOI: 10.1080/0142569900110405 (cit. on p. 14).

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning". In: *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019.* IEEE, 2019, pp. 739–753. DOI: 10.1109/SP.2019.00065 (cit. on p. 24).

[Nat02] National Center for Education Statistics, ed. *Education Longitudinal Study of 2002.* 2002 (cit. on p. 53).

[Nea+18] Tempestt J. Neal et al. "Surveying Stylometry Techniques and Applications". In: *ACM Comput. Surv.* 50.6 (2018), 86:1–86:36. DOI: 10.1145/3132039 (cit. on p. 40).

[Och19] Rodrigo Ochigame. "The Invention of "Ethical AI"". In: *The Intercept* 20 (2019) (cit. on pp. 4, 13).

[Och20] Rodrigo Ochigame. "The Long History of Algorithmic Fairness". In: *Phenomenal World* (Jan. 2020) (cit. on pp. 6, 14).

[Opi+17] Marc Opijnen et al. "On-Line Publication of Court Decisions in the EU: Report of the Policy Group of the Project "Building on the European Case Law Identifier"". In: *Available at SSRN 3088495* (2017) (cit. on pp. 31, 36).

[Org11] Organisation for Economic Co-operation and Development. *The Call for Innovative and Open Government: An Overview of Country Initiatives*. Organisation for Economic Co-operation and Development, 2011. ISBN: 978-92-64-10704-5 (cit. on p. 30).

[Pap+17] Nicolas Papernot et al. "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017 (cit. on p. 49).

[Pea19] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Commun. ACM* 62.3 (2019), pp. 54–60. DOI: 10.1145/3241036 (cit. on p. 27).

[PRT08] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. Ed. by Ying Li, Bing Liu, and Sunita Sarawagi. ACM, 2008, pp. 560–568. DOI: 10.1145/1401890.1401959 (cit. on p. 26).

[PLP04] Luc Plamondon, Guy Lapalme, and Frédéric Pelletier. "Anonymisation de Décisions de Justice." In: *XIe Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN 2004)*. Fès, Maroc: Bernard Bel et Isabelle Martin. (éditeurs) / Bernard Bel et Isabelle Martin. (éditeurs), May 2004, pp. 367–376 (cit. on p. 36).

[PB22] Drago Plecko and Elias Bareinboim. "Causal Fairness Analysis". In: *CoRR* (2022). DOI: 10.48550/arXiv.2207.11385. arXiv: 2207.11385 (cit. on pp. 16, 26).

[Ple+17] Geoff Pleiss et al. "On Fairness and Calibration". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 5680–5689 (cit. on p. 27).

[PPC16] Sabrina Praduroux, de Valeria Paiva, and di Luigi Caro. "Legal Tech Start-Ups: State of the Art and Trends". In: *Proceedings of the Workshop on MIning and REasoning with Legal Texts Collocated at the 29th International Conference on Legal Knowledge and Information Systems*. 2016 (cit. on p. 33).

[PB20] Justin P. Preddie and Monica Biernat. “More than the Sum of Its Parts: Intersections of Sexual Orientation and Race as They Influence Perceptions of Group Similarity and Stereotype Content”. In: *Sex Roles* 84 (2020), pp. 554–573 (cit. on pp. 5, 17).

[Pre23] President and Fellows of Harvard College, ed. *Caselaw Access Project*. 2023 (cit. on p. 33).

[Puj+20] David Pujol et al. “Fair decision making using privacy-protected data”. In: *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt et al. ACM, 2020, pp. 189–199. DOI: 10.1145/3351095.3372872 (cit. on pp. 28, 29).

[QG10] Paulo Quaresma and Teresa Gonçalves. “Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents”. In: *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*. Ed. by Enrico Francesconi et al. Vol. 6036. Lecture Notes in Computer Science. Springer, 2010, pp. 44–59. DOI: 10.1007/978-3-642-12837-0_3 (cit. on p. 34).

[Rat+20] Alexander Ratner et al. “Snorkel: rapid training data creation with weak supervision”. In: *VLDB J.* 29.2-3 (2020), pp. 709–730. DOI: 10.1007/s00778-019-00552-1 (cit. on p. 38).

[Rea+16] Sean Reardon et al. “Agent-Based Simulation Models of the College Sorting Process”. In: *J. Artif. Soc. Soc. Simul.* 19.1 (2016). DOI: 10.18564/jasss.2993 (cit. on pp. 58, 60).

[Rea+18] Sean F Reardon et al. “What Levels of Racial Diversity Can Be Achieved with Socioeconomic-Based Affirmative Action? Evidence from a Simulation Model”. In: *Journal of Policy Analysis and Management* (2018) (cit. on pp. 9, 52, 53, 55, 57–61, 95, 98, 100).

[Rea+14] Sean F. Reardon et al. *Agent-Based Simulation Models of the College Sorting Process*. Version 1.0.0. May 23, 2014 (cit. on p. 58).

[Rea+17] Sean F. Reardon et al. “What Levels of Racial Diversity Can Be Achieved with Socioeconomic-Based Affirmative Action? Evidence from a Simulation Model (CEPA Working Paper No.15-04)”. Dec. 2017 (cit. on pp. 61–67, 101, 102).

[Rid+21] Diane Ridgeway et al. *Challenge Design and Lessons Learned from the 2018 Differential Privacy Challenges*. NIST technical note 2151. National Institute of Standards and Technology, Apr. 2021. DOI: 10.6028/NIST.TN.2151 (cit. on p. 44).

[RG20] Maria Rigaki and Sebastian Garcia. “A Survey of Privacy Attacks in Machine Learning”. In: *CoRR* (2020). arXiv: 2007.07646 (cit. on p. 24).

[Rou+22] Gilles Rouet et al. *Algorithmes et Décisions Publiques*. CNRS Éditions via OpenEdition, 2022. DOI: 10.4000/lectures.36133 (cit. on p. 14).

[Rud18] Cynthia Rudin. "Please Stop Explaining Black Box Models for High Stakes Decisions". In: *CoRR* (2018). arXiv: 1811.10154 (cit. on p. 15).

[SBV13] David Sánchez, Montserrat Batet, and Alexandre Viejo. "Automatic General-Purpose Sanitization of Textual Documents". In: *IEEE Trans. Inf. Forensics Secur.* 8.6 (2013), pp. 853–862. DOI: 10.1109/TIFS.2013.2239641 (cit. on p. 41).

[SD87] NJ Scheers and C Mitchell Dayton. "Improved Estimation of Academic Cheating Behavior Using the Randomized Response Technique". In: *Research in Higher Education* 26 (1987), pp. 61–69. DOI: 10.1007/bf00991933 (cit. on p. 23).

[Sch+18] K. Schulte et al. "Analysis of the Eurotransplant Kidney Allocation Algorithm: How Should We Balance Utility and Equity?" In: *Transplantation Proceedings* 50.10 (Dec. 2018), pp. 3010–3016. DOI: 10.1016/j.transproceed.2018.08.040 (cit. on pp. 4, 13).

[Sho+17] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 3–18. DOI: 10.1109/SP.2017.41 (cit. on pp. 24, 25).

[Sim18] Ric Simmons. "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System". In: *UC Davis L. Rev.* 52 (2018), p. 1067. DOI: 10.2139/ssrn.3156510 (cit. on p. 14).

[SR20] Congzheng Song and Ananth Raghunathan. "Information Leakage in Embedding Models". In: *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*. Ed. by Jay Ligatti et al. ACM, 2020, pp. 377–390. DOI: 10.1145/3372297.3417270 (cit. on p. 24).

[SOQ18] SOQUIJ, ed. *Protection de La Jeunesse – 186470, 2018 QCCQ 6920*. 2018 (cit. on p. 39).

[SOT22] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic Data - Anonymisation Groundhog Day". In: *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*. Ed. by Kevin R. B. Butler and Kurt Thomas. USENIX Association, 2022, pp. 1451–1468 (cit. on pp. 8, 45).

[Swe96] Latanya Sweeney. "Replacing Personally-Identifying Information in Medical Records, the Scrub System." In: *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1996, p. 333 (cit. on pp. 37, 40).

[Swe00] Latanya Sweeney. “Simple Demographics Often Identify People Uniquely”. 2000 (cit. on p. 20).

[Swe02] Latanya Sweeney. “k-Anonymity: A Model for Protecting Privacy”. In: *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10.5 (2002), pp. 557–570. DOI: 10.1142/S0218488502001648 (cit. on pp. 5, 15, 20, 35, 37).

[Tan+18] Sarah Tan et al. “Investigating Human + Machine Complementarity for Recidivism Predictions”. In: *CoRR* (2018). arXiv: 1808.09123 (cit. on p. 34).

[Tan+17] Jun Tang et al. “Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12”. In: *CoRR* (2017). arXiv: 1709.02753 (cit. on pp. 5, 16, 22, 23).

[The+21] Jens T. Theilen et al. “Feminist data protection: an introduction”. In: *Internet Policy Rev.* 10.4 (2021). DOI: 10.14763/2021.4.1609 (cit. on p. 21).

[TKA17] D. Thenmozhi, Kawshik Kannan, and Chandrabose Aravindan. “A Text Similarity Approach for Precedence Retrieval from Legal Documents”. In: *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017*. Ed. by Prasenjit Majumder et al. Vol. 2036. CEUR Workshop Proceedings. CEUR-WS.org, 2017, pp. 90–91 (cit. on p. 33).

[Tom21] Stavros Tombazos. “La Reproduction Capitaliste Chez Marx”. In: *Actuel Marx* 2 (2021), pp. 77–95. DOI: 10.3917/amx.070.0077 (cit. on p. 14).

[Tra+16] Florian Tramèr et al. “Stealing Machine Learning Models via Prediction APIs”. In: *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016*. Ed. by Thorsten Holz and Stefan Savage. USENIX Association, 2016, pp. 601–618 (cit. on p. 24).

[Tyr00] Alan Tyree. “James Popple, A Pragmatic Legal Expert System. Applied Legal Philosophy Series”. In: *Artif. Intell. Law* 8.1 (2000), pp. 67–74. DOI: 10.1023/A:1008342900169 (cit. on p. 14).

[Uni09] University of Houston Law Center. *How to Brief a Case*. Tech. rep. University of Houston Law Center, 2009 (cit. on p. 32).

[VK22] Thomas F. Varley and Patrick Kaminski. “Untangling Synergistic Effects of Intersecting Social Identities with Partial Information Decomposition”. In: *Entropy* 24.10 (2022), p. 1387. DOI: 10.3390/e24101387 (cit. on pp. 5, 17).

[WCV11] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. “The NumPy Array: A Structure for Efficient Numerical Computation”. In: *Comput. Sci. Eng.* 13.2 (2011), pp. 22–30. DOI: 10.1109/MCSE.2011.37 (cit. on p. 58).

[War65] Stanley L. Warner. “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias.” In: *Journal of the American Statistical Association* (1965), pp. 63–66. DOI: 10.1080/01621459.1965.10480775 (cit. on p. 23).

[WK18] Benjamin Weggenmann and Florian Kerschbaum. "SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. Ed. by Kevyn Collins-Thompson et al. ACM, 2018, pp. 305–314. DOI: 10.1145/3209978.3210008 (cit. on pp. 40, 42).

[Wor02] World Legal Information Institute, ed. *Declaration on Free Access to Law*. 2002 (cit. on p. 30).

[Xu+21] Zhen Xu et al. "Codabench: Flexible, Easy-to-Use and Reproducible Benchmarking for Everyone". In: *CoRR* (2021). arXiv: 2110.05802 (cit. on p. 45).

[Ye+22] Jiayuan Ye et al. "Enhanced Membership Inference Attacks against Machine Learning Models". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*. Ed. by Heng Yin et al. ACM, 2022, pp. 3093–3106. DOI: 10.1145/3548606.3560675 (cit. on p. 25).

[Yeo+18] Samuel Yeom et al. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*. IEEE Computer Society, 2018, pp. 268–282. DOI: 10.1109/CSF.2018.00027 (cit. on pp. 25, 49).

[ZYS23] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in Ranking, Part I: Score-Based Ranking". In: *ACM Comput. Surv.* 55.6 (2023), 118:1–118:36. DOI: 10.1145/3533379 (cit. on p. 69).

[Zem+13] Richard S. Zemel et al. "Learning Fair Representations". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 325–333 (cit. on p. 27).

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*. Ed. by Jason Furman et al. ACM, 2018, pp. 335–340. DOI: 10.1145/3278721.3278779 (cit. on p. 27).

[Zha+18] Dan Zhang et al. "EKTELO: A Framework for Defining Differentially-Private Computations". In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*. Ed. by Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein. ACM, 2018, pp. 115–130. DOI: 10.1145/3183713.3196921 (cit. on p. 42).

[Zha+17] Jun Zhang et al. "PrivBayes: Private Data Release via Bayesian Networks". In: *ACM Trans. Database Syst.* 42.4 (2017), 25:1–25:41. DOI: 10.1145/3134428 (cit. on p. 49).

[ZZL15] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. "Character-level Convolutional Networks for Text Classification". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Corinna Cortes et al. 2015, pp. 649–657 (cit. on p. 34).

[Zha+20] Yuheng Zhang et al. "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 250–258. DOI: 10.1109/CVPR42600.2020.00033 (cit. on p. 24).

[Zhe+18] Jianming Zheng et al. "A Hierarchical Neural-Network-Based Document Representation Approach for Text Classification". In: *Mathematical Problems in Engineering* 2018 (2018). DOI: 10.1155/2018/7987691 (cit. on p. 34).

[Zho+20] Jianlong Zhou et al. "A Survey on Ethical Principles of AI and Implementations". In: *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, Canberra, Australia, December 1-4, 2020*. IEEE, 2020, pp. 3010–3017. DOI: 10.1109/SSCI47803.2020.9308437 (cit. on pp. 4, 13).

A. College-Student model

In this appendix we present in details the model as described in [Rea+18].

A.1. Notation and variables

Variables All used variables along with their dimensions and indexes are listed in Table A.1. We note S the set of students, C the set of colleges, and Y the set of iterations.

A.2. Model

For each run of the model, $J = 40$ colleges are generated. At each iteration, a new cohort of $N = 10000$ students is sampled. Colleges are represented by a quality attribute Q, initially drawn from a normal distribution $\mathcal{N}(130, 1100)$. Students have three attributes: race, resources R, and achievement A. The latter two are sampled from race-specific bivariate normal distributions with non-zero correlation.

Three successive decisions are taken at each iteration:

1. Students *apply* to a subset of colleges;
2. Colleges *admit* a subset of students from the applications;
3. Students *enroll* into the best college they are admitted to.

Race is samples from a categorical distribution as given in Table A.2.

Resources and achievement are sampled from bivariate normal distributions with non-zero correlation and conditioned on race. The student model can be expressed as the following mixture model, with parameters of the components given in Table A.3.

$$\mathsf{Race} \sim \text{Categorical}(\theta_{\mathsf{Race}}) \tag{A.1}$$

$$\mathsf{Res}, \mathsf{Ach} \mid \mathsf{Race} = Normal_2(\mu_z, \Sigma_z) \tag{A.2}$$

Figure 5.1 proposes an abstract view of the model, dropping the educational context. Individuals have a sensitive attribute Z, which impacts their attributes I and E. Attributes I is innate–that is they cannot be acted upon–and attributes E are external (i.e. received). Those two kinds of attributes are combined as an observed feature set X, with the "objective" of X capturing I but suffering the impact of E. A decision Y is taken on X, and on some context Q independent of the individuals. The decision then impacts future contexts and future external variables.

Name	Index	Dimension	Description
Race	y, s	$Y \times S$	Student's race
Res	y, s	$Y \times S$	Student's resources
Ach	y, s	$Y \times S$	Student's achievement
Qual	y, c	$Y \times C$	College's quality
SAch	y, s	$Y \times S$	Student's perception of self achievement Ach
CAch	y, c, s	$Y \times C \times S$	College's perception of a student's achievement Ach
SQual	y, c, s	$Y \times C \times S$	Student's perception of a college's quality Qual
NApp	y, s	$Y \times S$	Student's number of applications
NAdm	y, c	$Y \times C$	College's number of admissions
YieldEnr)	y, c	$Y \times C$	College's enrollment yield
ρ_S^{A}	y, s	$Y \times S$	Reliability of SAch
ρ_S^{Q}	y, s	$Y \times S$	Reliability of SQual
ρ_C^{A}			Reliability of CAch
Util	y, c, s	$Y \times C \times S$	Student's perceived utility of attenting a college
PAdm	y, c, s	$Y \times C \times S$	Probability of a student being admited to a college
App	y, c, s	$Y \times C \times S$	Application of student to college
Adm	y, c, s	$Y \times C \times S$	Admission of student to college
Enr	y, c, s	$Y \times C \times S$	Enrollment of student to college

Table A.1.: Variables

Code	Name	Probability
A	Asian	0.05
B	Black	0.15
H	Hispanic	0.2
W	White	0.6

Table A.2.: Race distribution

Race	Resources		Achievement		Correlation ρ
	μ_R	σ_R	μ_A	σ_A	
Asian	0.012	0.833	1038	202	0.441
Black	-0.224	0.666	869	169	0.305
Hispanic	-0.447	0.691	895	185	0.373
White	0.198	0.657	1052	186	0.395

Table A.3.: Parameters of ressources and achievement

A.2.1. Application

During the first stage, students select a subset–called portfolio–of colleges to apply to. Students observe colleges' quality with some amount of uncertainty which represents "imperfect information and idiosyncratic preferences". This error depends on students' resources as a way to model high-resources families having better information about college quality.

$$\rho_S^{\mathsf{Q}} = \max(0.5, \min(0.9, 0.7 + 0.1 \times \mathsf{Res}) \tag{A.3}$$
$$\tau = \mathrm{var}(\mathsf{Qual}) \times (1 - \rho_S^{\mathsf{Q}}) \times 1/\rho_S^{\mathsf{Q}} \tag{A.4}$$
$$u \sim N(0, \tau) \tag{A.5}$$
$$\mathsf{SQual} = \mathsf{Qual} + u \tag{A.6}$$

Students use observed college quality to evaluate the utility of attending that college:

$$\mathsf{Util} = \mathsf{SQual} - 250 \tag{A.7}$$

Students may augment their own achievement, and they perceive their own achievement with noise. Thus, their assessment of their achievement, for purposes of deciding where to apply, is:

$$\rho_S^{\mathsf{A}} = \max(0.5, \min(0.9, 0.7 + 0.1 \times \mathsf{Res}) \tag{A.8}$$
$$\sigma = \mathrm{var}(A)1 - \rho_S^{\mathsf{A}} \times 1/\rho_S^{\mathsf{A}} \tag{A.9}$$
$$e \sim N(0, \sigma) \tag{A.10}$$
$$\alpha = 0.1 \times \mathsf{Res} \times \sigma_{\mathsf{Ach,Race}} \tag{A.11}$$
$$\mathsf{SAch} = \mathsf{Ach} + \alpha + e \tag{A.12}$$

where α_s represents enhancements such as extracurricular activities. As above, the error in a student's assessment of his or her own achievement has a variance that depends on his or her family resources.

Based on their noisy observations of their own achievement and college quality, students estimate their probabilities of admission into each college using a logistic model fitted on admission patterns over the prior 5 years. For the first 5 years of the simulation, the intercept is set to 0 and the slope to -0.015.

Each student applies to a set of $\mathsf{NApp} = 4 + [0.5 \times R]$ colleges that maximize their overall expected utility.

A.2.2. Admission

The second decision phase consists in the admission of students by colleges.

Colleges observe the achievement of applicants with some noise, similar to the observation of quality by students and serves the same purpose. Colleges assess students'

achievement with a reliability of 0.8.

$$\phi = \text{var}(A) \times \frac{1 - 0.8}{0.8} \tag{A.13}$$

$$w \sim N(0, \phi) \tag{A.14}$$

$$A^{**} = A_s + \alpha + w \tag{A.15}$$

"in the model, colleges' uncertainty and idiosyncratic preferences have the effect of adding noise with a standard deviation of 100 points (half a standard deviation of achievement) to each student's application."

Colleges rank applicants according to A^{**} and admit the top $NAdm = 150/Yield$ applicants with the objective of enrolling 150 students. The first iteration, colleges assume a proportion of admitted students expected to enroll *yield* computed as

$$Yield = 0.2 + 0.6 \times (\texttt{College rank percentile}) \tag{A.16}$$

"The lowest-quality college expects slightly over 20 percent of admitted students to enroll and the highest quality college expects 80 percent of admitted students to enroll." In subsequent iterations, colleges use up to 3 years of enrollment history to compute the expected yield.

Note that, alternatively, we can use the survival function of the Qualdistribution to model the college rank percentile.

A.2.3. Enrollment

Students enroll in the college with the highest estimated utility of attendance (U^*) to which they were admitted.

A.2.4. Iteration

Colleges' quality (Q) are updated each year based on the new enrolled students as:

$$Q' = 0.09 \times Q + 0.1 \times \bar{A}, \tag{A.17}$$

where $\bar{A}$ is the average value of A among the newest cohort of students enrolled in college c.

A.2.5. Simulation duration

The scenarios in [Rea+18] last 30 or 50 iterations, with the first 10 years used as "bootstrapping" to stabilizes iteration-dependent variables such as the logistic regression or colleges' quality.

B. Further questions of reproductibility

Condition	None	Applied	Admitted	**Original**
Mean achievement	990.01	979.33	1177.59	~ 800
Prop. minority	0.35	0.36	0.17	~ 0.6
Prop. low-income	0.6	0.59	0.4	~ 0.5

Table B.1.: Values for Figure 5.9 (red arrow only) when considering students who do not enroll without condition, while having applied, or while being admitted. The last row displays the values from the original figures. Restricting to student having applied gives similar results as having no condition, while capturing only students with an admission gives values further to the original ones.

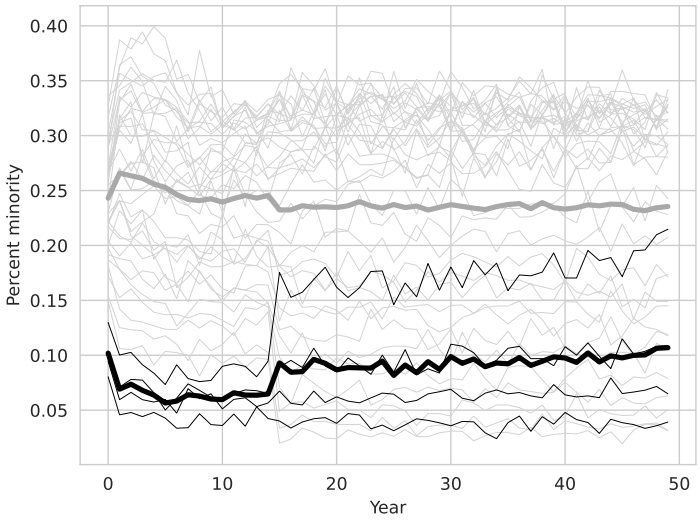


Figure B.1.: Minority enrollment with the top four colleges using strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100) *using the formula provided in [Rea+18] for the SES-based affirmative action.* See Section 5.4

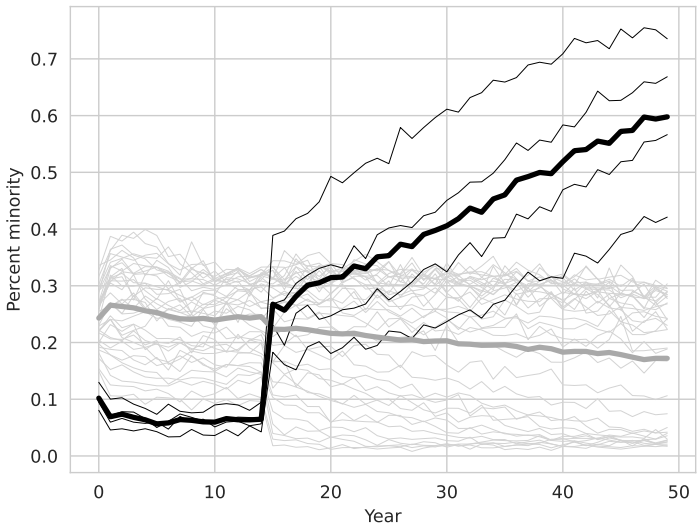


Figure B.2.: Minority enrollment with the top four colleges using strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100) *with penalization of high-income students (no truncation to* ≥ 0*).* See Section 5.4

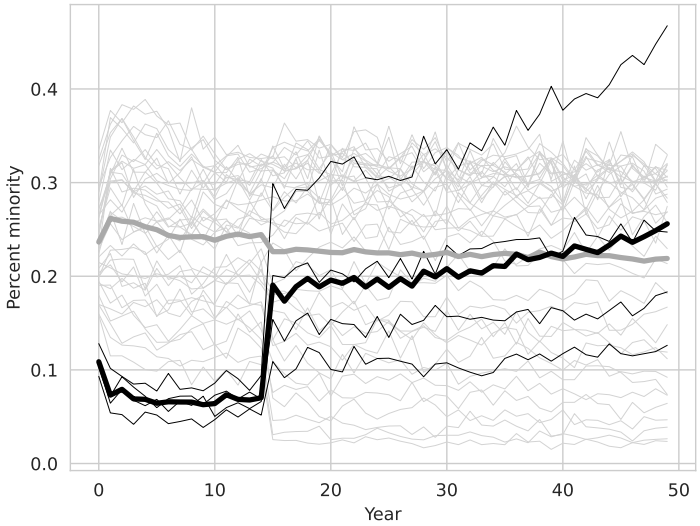


Figure B.3.: Minority enrollment with the top four colleges using strong SES-based affirmative action (weight of 150) and strong race-based recruitment (weight of 100) *truncating reliability of student perceptions of college quality to 0.7.* See Section 5.4

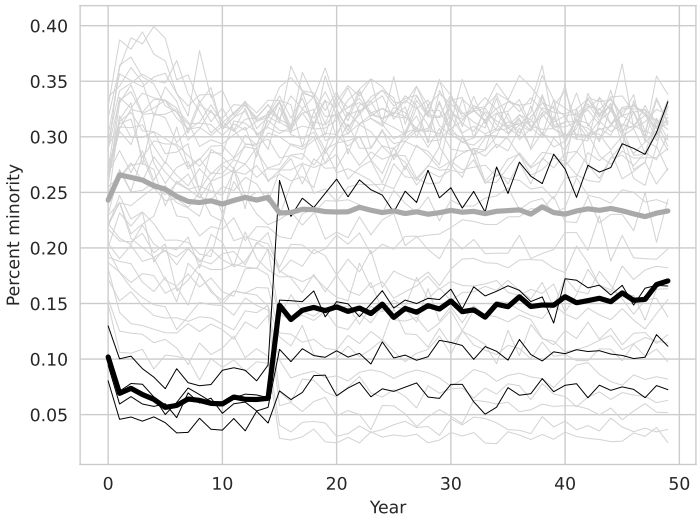


Figure B.4.: "Changes in Black and Hispanic Enrollment over Time with Top 4 Colleges using Moderate SES-Based Affirmative Action and Moderate Race-Based Recruiting" [Rea+17, Figure C3]

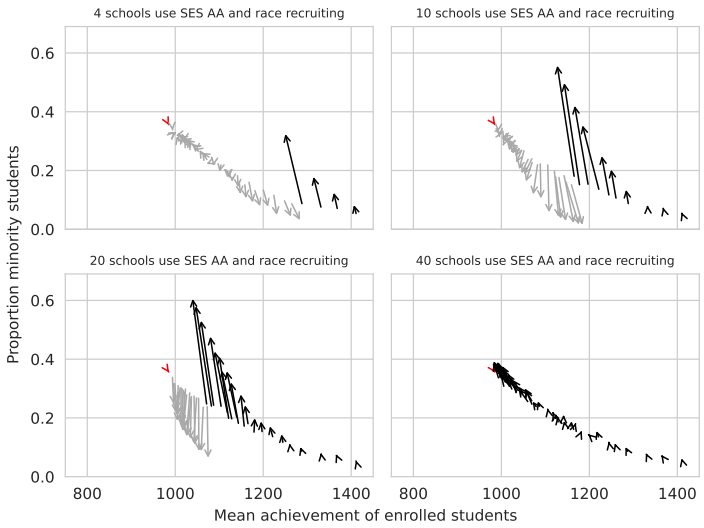


Figure B.5.: "The mean achievement and proportion minority by number of schools using admissions policies." [Rea+17, Figure D2]

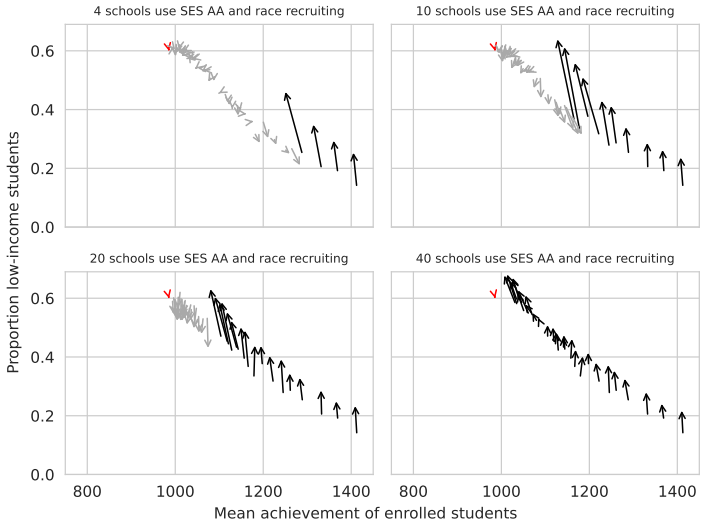


Figure B.6.: "The mean achievement and proportion low-income by number of schools using affirmative action." [Rea+17, Figure D3]

COLLEGE DOCTORAL BRETAGNE | MATHS, TELECOMS INFORMATIQUE, SIGNAL SYSTEMES, ELECTRONIQUE

Université de Rennes

UQÀM

Titre : Enjeux de vie privée et d'équité posés par les systèmes de décision algorithmiques

Mot clés : vie privée, équité, système de décision, socio-technique

Résumé : Les algorithmes sont de plus en plus utilisés pour prendre des décisions impactant les individus, les cohortes et la société dans son ensemble. Cette ubiquité soulève d'importantes préoccupations éthiques et sociales, notamment la protection de la vie privée et l'équité. Cette thèse étudie ces deux sujets techniques sous l'angle de leur utilisation pratique et de leurs exigences sociétales. Dans une première contribution, nous examinons le conflit entre la vie privée et la transparence lors de la publication de procédures judiciaires. Dans une deuxième contribution, nous proposons un framework pour organiser des compétitions d'attaque de mécanismes de protection de la vie privée afin de mieux établir leur comportement dans la pratique. Dans une troisième contribution, nous nous concentrons sur les limites des définitions techniques d'équité et utilisons une simulation ancrée dans la réalité comme moyen d'observer leur impact à long terme sur l'ensemble d'un système. L'objectif principal de cette thèse est de mettre en évidence les questions fondamentales soulevées par la technicisation des défis sociaux ainsi que de proposer des outils techniques et des analyses visant à ramener ces défis sociotechniques dans leur contexte social.

Title: Privacy and fairness issues with algorithmic decision systems

Keywords: privacy, fairness, decision systems, sociotechnical

Abstract: Algorithms are increasingly used across all layers of society, including in high-stake decision systems, impacting individuals, cohorts and society as a whole. This omnipresence of algorithms raises important ethical and social concerns, including in particular privacy and fairness. In this thesis, we study these two technical subjects under the lens of their practical usage and strong societal requirements. In a first contribution we investigate the conflict between privacy and transparency when publishing legal proceedings. In a second contribution, we propose a framework to organize privacy challenges with a focus on attacking privacy-preserving data publishing mechanisms to better define their behavior in practice. As a third contribution, we focus on the limits of technical fairness definitions and leverage a simulation grounded in reality as a way to observe the long-term impact of fairness on a whole system. Overall, the main objective of this thesis is to highlight fundamental issues raised by the technicization of social challenges as well as to propose technical tools and analyses aimed towards bringing these sociotechnical challenges back to their social context.