Digital Discrimination[1]

Natalia Criado, Jose M. Such
Department of Informatics
King's College London
United Kingdom
{natalia.criado,jose.such}@kcl.ac.uk

Abstract

In digital discrimination, users are treated unfairly, unethically or just differently based on their personal data that is automatically processed by an algorithm. Digital discrimination very often reproduces the existing instances of discrimination in the offline world by either inheriting the biases of prior decision makers, or simply reflecting widespread prejudices in society. It may also exacerbate existing inequalities by causing less favourable treatment for historically disadvantaged groups. This chapter analyses this challenging problem, the existing solutions and their limitations from a legal and computer science prespective. The chapter concludes by proposing a roadmap of open issues to overcome these limitations.

Index terms: Digital Discrimination, Bias, Machine Learning, Artificial Intelligence

1 Introduction

Digital discrimination is a form of discrimination in which automated decisions taken by algorithms, increasingly based on Artificial Intelligence techniques like Machine Learning, treat users unfairly, unethically or just differently based on their personal data (Such, 2017) such as income, education, gender, age, ethnicity, religion. Digital discrimination is becoming a serious problem (O'Neil, 2016), as more and more tasks are delegated to computers, mobile devices, and autonomous systems, e.g., some UK firms base their hiring decisions on automated algorithms[2].

In this chapter, we briefly introduce the notion of unlawful discrimination based on current legislation in selected jurisdictions and consider how that notion is currently extended to the digital world in the form of digital discrimination. Then, we show some of the most relevant digital discrimination examples studied in the literature. We also review advances in research advances in seeking to address digital discrimination and outline a research roadmap to tackle the challenges to detect and address digital discrimination.

1.1 Discrimination

The majority of nations have legislation prohibiting discrimination; e.g., the International Covenant on Civil and Political Rights, the U.S. Civil Rights Act, the European Convention for the Protection of Human Rights, the UK Equality Act 2010, and so on and so forth. However, there is not a universally accepted definition of discrimination, what it is and when it does

[1] This is the authors' version of a chapter to appear in "Algorithmic Regulation", Oxford University Press, 2019.

[2] http://www.bbc.co.uk/news/business-36129046

happen. Indeed, it is a concept very much shaped by culture, social and ethical perceptions, and historical and temporal considerations. Most anti-discrimination legislation simply consists of a non-exhaustive list of criteria or protected attributes (e.g., race, gender, sexual orientation) on the basis of which discrimination is forbidden. Thus, discrimination are actions, procedures, etc., that disadvantage citizens based on their membership to particular social groups defined by those attributes. Legal systems traditionally distinguish between two main types of discrimination (Altman, 2016):

- Direct Discrimination (also known as Disparate Treatment) considers the situations in which an individual is treated differently because of their membership to a particular social group. This ultimately means that different social groups are being treated differently, with some of them effectively being disadvantaged by these differences in treatment. A clear example of direct discrimination would be a company having the policy of not hiring women with young children. Note, however, that direct discrimination does not necessarily involve intentionality or that the discrimination process is explicit. In particular, direct discrimination is a more complex phenomena that can take many shapes regardless of discrimination being explicit or intentional. There are therefore two dimensions of discrimination outlined below.
    - Explicit/Implicit. Direct Discrimination can be explicit, exemplified in the previous case of member of a particular social group (women with young children) is explicitly disadvantaged by a decision process (hiring policy). More subtle cases of direct discrimination can also occur in which the disadvantaged group is not explicitly mentioned. For example, the same company may replace the explicit hiring policy with a policy of not hiring candidates who have had a career break in recent years. The new policy does not explicit refer to the relevant social group. Instead, it employs some facially-neutral criteria that accomplishes the same discrimination aim. This is an example of implicit direct discrimination as the alternative policy was created with the aim of discriminating against women with young children (who are statistically more likely to have had a recent career break).
    - Intentional/Unintentional. Direct discrimination is not just intentional discrimination, for example a teacher who encourages male students to support and help the only female student in the class may be unintentionally discriminating by adopting a patronizing attitude towards that female student. In this situation, the offender agent may not be aware of the discriminatory motive behind their act, e.g., they may not be aware of the fact that they prejudice women as being weaker and in need of support as the reason for their act.
- Indirect Discrimination (also known as Disparate Impact) considers the situations in which an apparently neutral act has a disproportionate negative effect on the members of a particular social group. This is considered discrimination even if there is no intention to discriminate that particular group or if there is not any unconscious prejudice motivating the discriminatory act. For example, a company having the policy to only consider customer satisfaction scores to award pro- motions may have a disproportionate impact on women, as there is empirical evidence suggesting that women are under evaluated when compared to their male counterparts with a similar objective performance[3]. In this case, the company may not have the intention to discriminate female employees, but the promotion criteria set may effectively disadvantage them disproportionally.

[3] http://www.wordstream.com/blog/ws/2014/05/13/gender-bias

1.2 Digital Discrimination

The term digital discrimination (Wihbey, 2015) refers to those direct or indirect discriminatory acts that are based on the automatic decisions made by algorithms[4]. Increasingly, decisions (even very important ones) are delegated to algorithms, such as those that use artificial intelligence techniques such as machine learning. From the jobs we apply for, to the products we buy, to the news we read and to the persons we date, many sensitive decisions are increasingly delegated to or, at least, influenced by those systems. Machine learning is a subfield of artificial intelligence that focuses on the study of computer algorithms that improve automatically through experience. Machine learning algorithms are classified into unsupervised algorithms (e.g., those used in data mining) that find patterns in a given dataset; and supervised algorithms that are presented with example inputs and their desired outputs to learn a general rule that maps inputs to outputs. What is predicted by the supervised algorithm is known as the target variable. This target variable can be nominal, i.e., its value is known as the class category and the machine learning task is known as classification; or numeric, i.e., the machine learning task is known as regression. For example, an algorithm trying to predict political affiliation from social network data has a nominal target variable where the possible values are the different political parties, whereas an algorithm trying to predict income from purchase data has a numeric target variable.
The automated decisions made by algorithms, including those based on machine learning, are sometimes perceived as faultless, not having most of the shortcomings that we humans have (e.g., tiredness or personal prejudices); and their decisions may be less scrutinized, that is, decisions made by algorithms are less closely examined than the decisions made by humans (Angwin, Larson, Mattu, & Kirchner, 2016). However, automated decision-making, and in particular machine learning algorithms, are likely to inherit the prejudices of programmers, previous decisions, users and/or the society. This leads to discriminatory outcomes (Pasquale, 2015; O'Neil, 2016). Indeed, machine learning algorithms have the potential to discriminate more consistently and systematically and at a larger scale than traditional non-digital discriminatory practices.

The literature in the area of digital discrimination very often refers to the related concept of algorithmic bias (Danks & London, 2017). Despite playing an important role in digital discrimination, algorithm bias does not necessarily lead to digital discrimination per se, and this distinction between bias and actual discrimination is crucial:

1. In algorithmic terms, a bias means a deviation from the standard, but it does not necessarily entail a disadvantageous treatment to particular social groups. For example, an autonomous car that is biased towards safe driving decisions may deviate from the standard driving norms, but it is not discriminating users.
2. Algorithms need some sort of extent of bias, if the decision that an algorithm models is always aligned with the standard and/or it is random to some extent, then it makes little sense to use an algorithm to make that decision. Indeed, machine learning algorithms rely on the existence of some statistical patterns in the data used to train them, so that an algorithm can learn to predict or make the most suitable decision.

Therefore, while bias is a very useful concept and it is indeed related to digital discrimination, we focus in this chapter on the problematic instances of bias, and on the extent to which bias may lead to digital discrimination. There are different causes for digital discrimination which can be categorised into:

[4] Note the term digital discrimination has also been used to define traditional discrimination practices facilitated by online information (Edelman & Luca, 2014) or a discriminatory access to digital technologies or information (Weidmann, Benitez-Baleato, Hunziker, Glatz, & Dimitropoulos, 2016; Blansett, 2008), which we are not using here.

- **Modelling:** machine learning algorithms are usually used to make predictions or recommendations based on some data. Note for some problems there may be an objective and unambiguous definition of the target variable and its values; e.g., an algorithm trying to predict age from images, since there is a measurable and clear definition of age. For other problems the target variable and its categories are an artefact defined by the designer of the algorithm, e.g., the classification of potential employees into good/bad candidates is a socially constructed feature for which there is not a clear and unambiguous specification. The creation of these artificial variables and categories may lead to discrimination. For example, the definition of what makes a particular candidate suitable to be hired, and hence, the definition of that category in machine learning terms is based on subjective assessments of current and previous successful candidates. These subjective assessments are very likely to include prejudices. Even more, if by any chance the definition of what makes a particular candidate to be hired relates to some sort of personal or sensitive information (such as ethnicity) or proxies to this sensitive information (e.g., the postcode and income can be a good predictor of race), this can also lead to digital discrimination.

- **Training:** in addition to or despite of how the modelling has been done, machine learning algorithms learn to make decisions or predictions based on a dataset that contains past decisions. This dataset may be provided to the machine learning algorithm offline or online:
    - In *offline* training the machine learning algorithm undergoes a learning phase where a prediction model is build based on the information on the training dataset. Obviously, the system will likely learn to make discriminatory decisions if that dataset reflects existing prejudices of decision makers (e.g., only candidates from a particular group are identified as appointable candidates), under-represents a particular social group (e.g., a dataset that does not contain information about a particular social group is likely to lead to inaccurate decisions for that social group), over-represents a particular social group (e.g., usually people who do not adhere to stereotypes for a particular profession are disproportionately supervised and their mistakes and faults will be detected at higher rates), or reflects social biases (e.g., particular social groups will have less opportunities to obtain certain qualifications).
    - In *online* training the machine learning algorithms learn as they are used. For example, reinforcement learning algorithms (Kaelbling, Littman, & Moore, 1996) are rewarded for good decisions and punished for bad ones. This type of systems are not free from discrimination. In some cases, the data collected online may be unrepresentative (e.g., some disadvantaged social groups may be excluded from interacting with the data collection system), in other cases the data is representative but discriminative (e.g., women, which are more likely to be specialised in low paid sectors, may click more frequently on adverts for low paid works, which may reinforce the algorithmic rule to suggest these types of job offers to women[5]). Similarly, machine learning algorithms have been demonstrated to inherit prejudices when being trained with online text (Caliskan, Bryson, & Narayanan, 2017) or by interacting with users[6].

[5] https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/

[6] A prominent example that was featured extensively in the media was the case of the Microsoft Tay Chatbot (http://www.bbc.co.uk/news/technology-35890188). Tay was an machine learning chatbot that was designed to interact and engage with users on Twitter to "entertain and connect with others online through casual and playful conversation". It turned out that through interactions with other users, it learned from them in a bad way, and it went rogue and started swearing and making racist remarks and inflammatory political statements. Microsoft finally took Tay down and cancelled the project.

- **Usage:** a non-discriminatory machine learning algorithm can also lead to discrimination when it is used in a situation for which it was not intended. For example, an algorithm utilised to predict a particular outcome in a given population can lead to inaccurate results when applied to a different population, which may also disadvantage this population over others and cause discrimination. For example, a natural language processing tool that is trained on the data shared online by a given social group may not be able to process the online communications produced by another social group, even if both groups share the same language; e.g., research (Eisenstein, O'Connor, Smith, & Xing, 2014) suggests that different groups use different dialects and word-choices in social media.

2 Examples of Digital Discrimination

Digital discrimination has been mostly studied and demonstrated in the context of gender, race, and income, location and lifestyle. In the following, we briefly describe some of the most relevant discrimination examples studied in the literature.

Gender. Digital discrimination based on gender has been widely documented in reseach (Datta, Tschantz, & Datta, 2015; Kay, Matuszek, & Munson, 2015; Wagner, Garcia, Jadidi, & Strohmaier, 2015). All of these studies showed that machine learning algorithms have the potential to exacerbate and perpetuate gender stereotypes and increase gender segregation. In particular, Datta et al. (2015) demonstrated that Google showed males ads encouraging the use of coaching services for high paying jobs more frequently than females, which may lead to discriminate women and to increase the gender pay gap. Due to the opacity of the advertisement recommendation system, the authors could not determine the causes of this effect; it may be that the policy of the advertiser algorithm is to tailor the adverts shown based on gender, which is not illegal or discriminatory per se.; or that differences on the online behaviour showed by males and females has driven the algorithm to show these coaching services to man as they are more likely to click on them. Kay et al. (2015) studied gender representation in image search results for different professions. Their study showed a slight under-representation of women (when compared to actual gender distributions of the different professions considered in the study) and that the female gender for a given profession was usually depicted less professionally. Wagner et al. (2015) studied the representation of women and men in Wikipedia. Their findings suggest that women are well covered by Wikipedia, however there are significant differences in the way in which they are portrayed. Women pages contain more information about their personal lives and their pages are less central in the network of pages when compared to male pages. This ex- ample may not be considered as digital discrimination as the information on Wikipedia has been produced by human users, not an algorithm and it that sense it does not differ from traditional discrimination. However, when existing algorithms (e.g., search engines), use that information to compute their results they could discriminate women who can be under-represented (e.g., search algorithms use network centrality measures to rank the results of a given query).

Race or Ethnicity. Other well-known examples of digital discrimination are related to race discrimination (Sweeney, 2013; Angwin et al., 2016). Sweeney (2013) demonstrated that advertisements suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names, regardless of the existence of arrest records for those names. The causes of this discrimination are not clear: (i) it me be the case that advertisers have defined particular search terms as the target of their adverts (note targeting is critical to advertisement and not illegal); or (ii) it may be the case that user behaviour (e.g., clicks received by each advertisement) has driven the machine learning algorithms to make these suggestions with particular names.Angwin et al. (2016) demonstrated that one of the crime prediction algorithms most widely used in the US criminal justice system is not only inaccurate, but discriminative. For example, evidence shows that black subjects are more likely to be wrongly assessed with high risk of re-offending, whereas white subjects are more likely to be wrongly assessed with low risk of re-offending. Although race is not itself a feature used to generate this risk score, some of the features used (e.g., information about job status, family conviction antecedents) may be highly correlated with race, which may explain the accuracy difference across different races.

Income, Location & Lifestyle. Aspects related to income, location or lifestyle may also lead to digital discrimination. A very clear example of intentional direct discrimination is the current practice of targeting low-income population with high-interest loans[7]. Discrimination based on location, which may not be perceived as illegal or unfair, may also lead to discrimination based on income. For example, Valentino-Devries, Singer-Vine, and Soltani (2012) demonstrated that an online pricing algorithm considering the proximity of users to competitor's stores discriminated against low-income users who are more likely to live far from these competitor's stores. More subtle cases of digital discrimination are related to the under representation of some groups of people in datasets due to their income, location and/or lifestyle. For example, Lerman (2013) has reflected on the dangers of exclusion from data collection processes. As more and more decisions both in the private and public spheres are informed by data collected from citizens, there is an increasing risk of not considering particular disadvantaged groups who may not be able to participate in data collection processes; e.g., they may not have access to the technology that is used to collect the data. This has the potential to not only ignore the needs and views of these marginalised groups in critical policy-making and industry-related decisions about housing, health care, education, and so on; but also to perpetuate the existing disadvantages.

3 Research on Addressing Digital Discrimination

Previous machine learning research has focused on two main streams: detecting discrimination and avoiding discrimination. In the following we review the main works in these 2 areas.

3.1 Detecting Discrimination

The first line of defence against discrimination is the development of metrics and processes to detect discrimination. Discrimination detection metrics can be also used in developing and implementing techniques aimed at avoiding discrimination in selecting optimisation criteria when preprocessing datasets or training algorithms. Within this line of research, Zliobaite (2015) surveys different metrics that have been proposed to measure indirect discrimination in data and the decisions made by algorithms. The study also discusses other traditional statistical measures that could be applied to measure discrimination. In particular, discrimination measures are classified by the authors into: statistical tests, which indicate the presence of discrimination at dataset level; absolute measures, which measure the magnitude of the discrimination present in a dataset; conditional measures, which capture the extent to which the differences between groups are due to protected attributes or other characteristics of individuals; and structural measures, which identify for each individual in the dataset if they are discriminated. Similarly, Tramer et al. (2017) proposed FairTest, a methodology and toolkit combining different metrics to detect what they call unwarranted association, a strong association be- tween the outputs of an machine learning algorithm and features defining a protected group. In (Datta, Sen, & Zick, 2016) the authors proposed quantitative metrics to determine the degree of influence of inputs on outputs of decision making systems. Their paper is not primarily intended to detect discrimination, but the measures they propose have the potential to increase transparency of decisions made by opaque machine learning algorithms, which, in turn, may provide useful information for the detection of discrimination.

All of these measures assume that the protected ground (i.e., the protected characteristics such as race or sex in which decisions cannot be based on; or the protected groups which cannot receive a disparate impact and what counts as disparate impact) are externally given, for example, by a law. However, as noted in Section 1.1 discrimination laws are not exhaustive

[7] https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

in terms of all potential instances of discrimination or the grounds that may lead to discrimination. Thus, the applicability of these detection measures is limited by the lack of a clear definition of a protected ground for a given problem.

3.2 Avoiding Discrimination

Works on avoiding discrimination can be further classified according to the way in which discrimination is prevented: modifying the problem model, preprocessing the data to be used to train the algorithm, and modifying the algorithm to include non-discrimination as a criterion to maximize together with prediction accuracy.

3.2.1 Problem Model

One of the first attempts to avoid discrimination in algorithms consisted in modifying how the problem at hand is modelled, so that the protected information is not available. For example, research has been conducted on the use of machine learning itself, but with the aim to learn new representations of a problem that minimise the inclusion of sensitive characteristics, while still allowing for good performance in the original prediction task (Edwards & Storkey, 2015; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Louizos, Swer- sky, Li, Welling, & Zemel, 2015). Although these approaches achieve a relatively good trade-off between prediction accuracy and non-discrimination, it has been shown that the representations learned are not completely free from sensitive information. Another problem with these approaches is that they cannot address classification tasks that are highly dependent on the sensitive characteristics. In such tasks there is a need for a better understanding of what counts as discrimination.

3.2.2 Data Preprocessing

These techniques are somewhat independent of the algorithm used to make predictions about the data, as they focus on modifying the dataset that is used for training algorithms — see Section 1.2 for an explanation of the different types of uses of datasets for training machine learning algorithms. This has the advantage that once the dataset has been modified to avoid discrimination, it can be reused to train other algorithms, even regardless of the type of algorithm. Discriminatory datasets normally have biases, so that when they are used for training machine learning algorithms, this leads to disfavouring particular users or groups of users. One example is that of unbalanced datasets, which contain more instances for particular types of users than others. The problem of unbalanced datasets (i.e., datasets that do not contain a realistic distribution of the data) is not exclusive to digital discrimination, within the more traditional machine learning and data mining areas different techniques have been proposed to deal with unbalanced datasets such as: oversampling, which replicates data elements that are under-represented; undersampling, which eliminates data elements in the over-represented class; and resampling, which consists in exchanging labels on data elements. Within the digital discrimination domain, the methods proposed consist on re-labelling the data elements to ensure fairness (Zhang, Berg, Maire, & Malik, 2006), which has similarities to the re-sampling methods aforementioned. Other proposals have particularly focused on removing disparate impact (or indirect discrimination) from datasets (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015). In particular, the authors propose a test to detect disparate impact based on the "80% rule" advocated by the US Equal Opportunity Employment Opportunity Commission[8].

Detecting the extent of biases that may lead to discrimination in other types of data such as unstructured data, hyperlinks, text, etc. is ever more complicated. For example, textual data

[8] The 80% rule states that the selection rate of a protected group should be at least 80% of the selection rate of the non-protected group. Note there is significant controversy about this rule: a recent memorandum from the US Equal Employment Opportunities Commission suggests that a more defensible standard would be based on comparing a company's hiring rate of a particular group with the rate that would occur if the company simply selected people at random.

may also encode existing biases in a more subtle way (e.g., with text that may reinforce gender stereotypes) and algorithms trained to make decisions based on this data will inherit such biases. One of the most common ways to process textual data such that it can be used by a machine learning algorithm are word embeddings (Goldberg & Levy, 2014), a framework to represent text as vectors. In (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016), the authors proposed a methodology to modify embeddings to remove discriminative associations between words that encode gender stereotypes. This allows the modification of the vectors (i.e., the dataset) used to feed the machine learning algorithms.

3.2.3 Algorithm Modification

This area of research, also known as fair algorithms or fair machine learning has significantly expanded over the last years, giving rise to modifications and extensions of existing machine learning and data mining algorithms, which specifically aim to prevent discriminative outcomes.

One of the first works on the definition of fair algorithms was made in (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). In this work, the authors propose a classification technique ensuring that similar individuals are treated similarly (regardless of their belonging to protected groups). This work assumes the existence of a task-specific metric determining the similarity between individuals. Note the creation of this metric is problematic, since there may be no unambiguous definition of what constitutes similarity (e.g., how similarities between job applicants can be defined), the definition of such metric can incorporate existing decision-making prejudices about similarity (e.g., not all candidates willing to work longer hours may achieve the same productivity); or, even worse, it can entail a disparate impact (or indirect discrimination) to particular groups (e.g., considering a qualification as a criteria may have a significant negative impact on a particular group of the population less able to obtain it, regardless of their capability to perform well in the job). On the other hand, the elicitation of a similarity metric, if feasible, can help detect and reveal existing prejudices and discrimination in decision making practices.

In (Calders, Karim, Kamiran, Ali, & Zhang, 2013) the authors proposed a method to control the effect of different attributes in linear regression models. A linear regression model tries to predict a numeric outcome based on a set of attributes. In this work, the authors proposed methods to ensure that the differences among the mean prediction in different groups, which are defined in terms of attribute values, are explained in terms of non-protected attributes and to ensure that the prediction errors for the different groups are also explained by non-protected attributes. In both cases, there is a need for a definition of protected and non-protected attributes. Note also that non-protected attributes may be highly correlated with protected ones and, therefore, become a proxy for discrimination.

Raff, Sylvester, and Mills (2017) proposed a method to minimize discrimination by tree-based algorithms, which are machine learning algorithms that provide a degree of interpretability, which also has a positive impact in decision transparency. In particular, their method allows for the protection of both nominal and numeric protected attributes, whereas most of the existing literature only focuses on nominal protected attributes.

4 Unresolved Issues surrounding Digital Discrimination

Despite the existing research on digital discrimination, some of which we summarised in the previous section, there are still several unresolved issues. One particular area of interest for us, and a very important step to understand and avoid digital discrimination, is how to attest digital discrimination, i.e., testing whether existing or new algorithms and/or datasets contain biases to the extent that make them discriminate users. While previous research on detecting digital discrimination can indeed measure and/or limit the extent of bias of an algorithm/dataset, the lingering questions needing urgent attention are: how much bias counts as digital discrimination? That is, how much bias is too much? There are different measures proposed to quantify bias and unwanted associations between user attributes/variables— as detailed in Section 3.1, but they are not sufficient to fully understand whether digital discrimination is present or not from a legal, ethical and/or socially-acceptable point of view. Therefore, what constitutes digital discrimination and how to translate that into automated methods that attest digital discrimination in datasets and algorithms is a very important and difficult to challenge. This requires cross-disciplinary collaboration enabling a synergistic and transformative social, legal, ethical, and technical approach, which will ultimately allow the design of ground-breaking methods to certify whether or not datasets and algorithms discriminate by automatically verifying computational non-discrimination norms. In particular, we discuss below four cross-disciplinary challenges towards this endeavour.

4.1 Socio-Economic and Cultural dimensions of Digital Discrimination

There is recent research towards the understanding of users' perceptions of digital discrimination, particularly with regards to concrete domains like online targeted advertising (Plane, Redmiles, Mazurek, & Tschantz, 2017) and criminal risk prediction (Grgic-Hlaca, Redmiles, Gummadi, & Weller, 2018). For instance, the second case focus on why people perceive certain features as fair or unfair to be used in algorithms. While this is indeed going in the right direction and it represents a very good start, more re- search is needed to produce a strong empirical base to understand the socio-economic and cultural dimensions of digital discrimination. This should entail not only survey-based studies, but also the definition of spaces for both co-creation, bringing together researchers and technical and non-technical users; and cross-disciplinarity, drawing on expertise from a diverse research team including experts in machine learning, human-computer interaction, law, ethics, philosophy, social sciences and so on. These spaces should bring the social, economic and cultural dimensions of digital discrimination by examining datasets, algorithms and processes. Techno-cultural approaches such as (Cote & Pybus, 2016) seem particularly suitable for this purpose.

4.2 Legal-Ethical Digital Discrimination Frameworks

As mentioned earlier in Section 1.1, there is legislation around the globe that already targets discrimination, though in some instances it is not completely specified when it comes to identifying the legally protected ground. In addition, there is no universally accepted definition of discrimination, in relation to both identifying what it is and when it occurs. Indeed, it is a concept very much shaped by culture, social and ethical perceptions, and time. Most anti-discrimination legislation sim- ply consists of a non-exhaustive list of criteria or protected attributes (e.g., race, gender, sexual orientation) on the basis of which discrimination is forbidden. Therefore, there is a need for the definition of a legal and ethical framework to articulate, define and describe digital discrimination. This entails the development of a new way of understanding discrimination from the perspective of decision-making algorithms. Based on a review of discrimination law and the socio-economic and cultural empirical base mentioned above, a critical reflection about the concept of discrimination and anti-discrimination within a digital context should be undertaken. The reflection should address issues such as: discrimination vs. freedom of choice, positive discrimination, and intersectionality. Such reflection should formulate an initial set of ethical norms covering legislative gaps or misconceptions. Argumentation processes, through cases and counter examples must be used to put into question and criticize these ethical norms in a systematic manner. Finally, this challenge also entails the interrogation of the assumptions under which non-discrimination norms are computable; i.e., under which circumstances an effective algorithm can determine whether or to what extent a decision-making algorithm or dataset abides by a non-discrimination norm.

4.3 Formal Non-Discrimination Computational Norms

Following the legal-ethical framework discussed above, a formal definition of norms that capture the requirements for non-discrimination should be defined. In particular, this challenge may consider research in the area of normative multi-agent systems (Criado, Argente, & Botti, 2011), which combines normative systems with intelligent systems, by considering norms that are inherently subjective and ambiguous. Norms and Normative systems have been extensively studied in recent years particularly as a way of limiting the autonomy of autonomous systems to adhere to acceptable behaviours. Norms are usually defined formally using deontic logic to state obligations, prohibitions, and permissions, but other modalities such as commitments and other formalizations such as soft constraints have also been considered. Norms could be used to define non-discriminatory behaviours, e.g., norms could define acceptable treatment of users, so that other non-acceptable behaviours would be prohibited. Norms have the added benefit that they can be used to govern and promote appropriate behaviours considering the whole socio-technical spectrum, from non-autonomous to autonomous algorithms and systems to human users (Singh, 2013), and they could be very useful as a common language for humans and machines, which could foster transparency and accountability in turn. The formal representation of non-discrimination norms will also pave the way to the development of novel verification methods to determine whether instances of discrimination occur in algorithms and datasets.

4.4 Automated Certification of Non-discrimination Norms

Following from the previous challenge, once it is necessary to formalise non-discrimination norms based on socio-economic, cultural, legal and ethical dimensions, the next step would be to design automated methods to verify and certify that datasets and algorithms satisfy the non-discrimination computational norms, identifying the causes of discrimination otherwise. This will support an integrated and multi-level view of discrimination, i.e., from the formal norm being violated, down to the instantiation of the norm, and to the specific rules, variables and/or associations causing discrimination. This challenge would entail both white-box and black-box scenarios. For white-box scenarios —i.e., algorithms the code of which can be inspected and comprehended (e.g., they can be expressed as decision trees) or datasets that are used to train algorithms are available — it is possible to take a model checking approach (non-discrimination norms are operationalised as formal properties) and/or mathematical approach (non- discrimination norms are operationalised as mathematical formulas defined over datasets). Note that the tools developed for white-box scenarios would be particularly useful to algorithm developers and data curators in businesses or the public sector to make sure algorithms/datasets will not discriminate before releasing them. For black-box scenarios — i.e., the cases in which the algorithm or the dataset used to feed or train the algorithm is not available for inspection (e.g., in auditing scenarios, because of IP reasons), two different cases must be considered: i) that the algorithm can be isolated and put into a testing environment (e.g., by an independent auditing party); and ii) that there are only limited opportunities to interact with an algorithm, with limited or non-direct observability of input-output sequences and limited control of inputs provided. The tools that would be developed for black-box scenarios would be particularly useful to regulatory, auditing and certification bodies, law enforcement, and NGOs to certify the presence (or the absence) of digital discrimination.

5 Conclusion

In this chapter, we introduced the problem of digital discrimination. Firstly, we briefly revisited how the term discrimination is used in the law, and how that might extend to the digital world. In particular, we reviewed some prominent examples of digital discrimination. After this, we reviewed the main research streams in digital discrimination, and outlined a roadmap to one such research streams, which is how to assess whether a particular algorithm or dataset discriminates users. Because of the very nature of digital discrimination, which encompasses social, cultural, legal, ethical, and technical aspects, we very much encourage cross-disciplinary research to address the challenges to detect and address digital discrimination, which would ultimately create the new transdisciplinary field of attesting digital discrimination.

Acknowledgements

We would like to thank the EPSRC for supporting this research under grant EP/R033188/1 “DADD: Discovering and Attesting Digital Discrimination”.

References

Altman, A. (2016). Discrimination. In E. N. Zalta (Ed.), The stanford encyclopedia of philosophy (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. ProPublica, May, 23.

Blansett, J. (2008). Digital discrimination. Library Journal , 133 (13), 26–29.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349–4357).

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Control- ling attribute effect in linear regression. In Data mining (icdm), 2013 IEEE 13th international conference on (pp. 71–80).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356 (6334), 183–186.

Cote, M., & Pybus, J. (2016). Simondon on datafication. a techno-cultural method. Digital Culture & Society , 2 (2), 75–92.

Criado, N., Argente, E., & Botti, V. (2011). Open issues for normative multi-agent systems. AI communications, 24 (3), 233–264.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In Proceedings of the 26th international joint conference on artificial intelligence (pp. 4691–4697).

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Security and privacy (sp), 2016 ieee symposium on (pp. 598–617).

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015 (1), 92–112.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in the- oretical computer science conference (pp. 214–226).

Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of airbnb. com.

Edwards, H., & Storkey, A. (2015). Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 .

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. PloS one, 9 (11), e113114.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining (pp. 259–268).

Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In WWW.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4 , 237– 285.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd annual acm conference on human factors in computing systems (pp. 3819–3828).

Lerman, J. (2013). Big data and its exclusions. Stan. L. Rev. Online, 66 , 55.

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. arXiv preprint arXiv:1511.00830 .

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Plane, A. C., Redmiles, E. M., Mazurek, M. L., & Tschantz, M. C. (2017). Exploring user perceptions of discrimination in online targeted advertising. In Usenix security.

Raff, E., Sylvester, J., & Mills, S. (2017). Fair forests: Regularized tree induction to minimize model bias. arXiv preprint arXiv:1712.08197 .

Singh, M. P. (2013). Norms as a basis for governing sociotechnical systems.
ACM TIST , 5 (1), 21.

Such, J. M. (2017). Privacy and autonomous systems. In Proceedings of the 26th international joint conference on artificial intelligence (pp. 4761–4767).

Sweeney, L. (2013). Discrimination in online ad delivery. Queue, 11 (3), 10.

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert,
M., . . . Lin, H. (2017). Fairtest: Discovering unwarranted associations in data-driven applications. In Security and privacy (euros&p), 2017 ieee european symposium on (pp. 401–416).

Valentino-Devries, J., Singer-Vine, J., & Soltani, A. (2012). Websites vary prices, deals based on users information. Wall Street Journal , 10 , 60–68.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In Icwsm (pp. 454–463).

Weidmann, N. B., Benitez-Baleato, S., Hunziker, P., Glatz, E., & Dimitropoulos, X. (2016). Digital discrimination: Political bias in internet service provision across ethnic groups. Science, 353 (6304), 1151–1155.

Wihbey, J. (2015). The possibilities of digital discrimination: Research on e-commerce, algorithms and big data. Journalists resource.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In International conference on machine learning (pp. 325–333).

Zhang, H., Berg, A. C., Maire, M., & Malik, J. (2006). Svm-knn: Discrim- inative nearest neighbor classification for visual category recognition. In Computer vision and pattern recognition, 2006 ieee computer society conference on (Vol. 2, pp. 2126–2136).

Zliobaite, I. (2015). A survey on measuring indirect discrimination in ma- chine learning. arXiv preprint arXiv:1511.00148 .