A roadmap for research in responsible data science

$Peter\ Bloem^*$ - Vrije Universiteit Amsterdam
$Oana\ Inel^*$ - Delft University of Technology
$Linda\ Rieswijk^*$ - Maastricht University

Motivation

Data and algorithms play an increasing role in all aspects of our daily lives. Algorithms help us decide which items to buy, what news articles to read, and which movies to watch. At an institutional level, the basic building blocks of our societal infrastructures are replaced and augmented by autonomous or semi-autonomous processes built on methods developed in data science. Criminal justice, healthcare provision, employment, and many other domains where the stakes are high, depend increasingly on analytics, machine learning, and statistics.

The impact can be far-reaching. When data-based autonomous decision making is implemented carelessly and non-transparently, serious damage can be done on a large scale. This became painfully clear in the Netherlands in 2020, when a parliamentary inquiry was held into false allegations of childcare benefits fraud made by the Tax and Customs administration. At least 26,000 parents in the Netherlands were found to be victims of wrong suspicions of fraud with the childcare allowance or were the victims of an overly tough approach to detect fraud by the tax authorities from 2013 to at least 2019.

A key aspect of the affair was the use of algorithmic decision making based on statistical patterns [AP]. One such system, the risk indicator, generated lists of candidate citizens to be checked for fraud. The features for this classification included, among other things, the nationality of the subject (Dutch/non-Dutch). The system was a complete black box, and investigators had no insight into why people appeared as having high risk. Currently, the common understanding of the system is that people with a risk level above 0.8 (out of 1) were automatically investigated, making the decision to investigate an autonomous one made by the system without human intervention.

These were AI methods implemented on a national scale and doing damage proportionally. Thousands of families were unjustly thrown into financial disarray for years. In response, several ministers resigned first, and finally, the Dutch government fell.

The conclusion is clear, careless, non-transparent use of data science (including data and algorithms) can amplify harmful institutional problems and serve to hide dangerous practices from scrutiny. The childcare benefits scandal clearly shows that there is a need to practice data science *responsibly*.

* Equal contribution, authors listed alphabetically

Furthermore, this raises many questions. How do we recognize irresponsible use of data science before it is too late? What does responsible data science entail? How can people with different backgrounds and values agree on a fundamental distinction between responsible and irresponsible data science?

In this document, we want to outline where the various academic fields that come under the umbrella of data science --- and some that do not yet --- can contribute to answering these and other questions. We propose a roadmap to highlight parts of the landscape of potential research that are particularly relevant to the questions of responsible data science. In data science, we identify two types of researchers; the ones that, just like the practitioners, do data science and the ones that study data science itself to produce methods and instruments that bring responsible data science easily and effectively. We hope that this may serve as inspiration to both types of researchers for future research projects, as a guide for funding agencies and policymakers on where to look for and to stimulate research output, and as a tool for practitioners, to see what aspects of data science practice must be considered if the methods are to be applied responsibly.

Data science is a hybrid field with practical applications. It borrows methods and concepts from various domains, such as statistics, knowledge representation, database, visualization, business science, machine learning, and information retrieval, among others. With this diverse mixture of methodologies, data science inherits several *risks*, such as using these methods socially, ethically, legally or technically irresponsibly. All these aspects urge us to discuss and draw certain boundaries for making the use of data and algorithms a responsible practice.

Much has been made in recent years of the social responsibilities around machine learning technology, particularly algorithmic bias. However, the use of data as a representation of the world is also bringing challenges. Data is always at the core of algorithms, being a crucial component for advancing and assessing the technological field. Nevertheless, irresponsible use of data can also create and foster inequality and inequity, perpetuate bias and prejudice, and even produce unlawful or unethical outcomes. Data science certainly inherited such problems, but they are only part of how data science can be misused. In this document, we hope to provide a comprehensive picture and discussion regarding the problems that arise from methods taken from, for instance, knowledge representation and business science.

In this document, we aim to structure the landscape and highlight some examples of fruitful and important research questions in this area, particularly those that may have been overlooked in existing work or that are not particularly streamlined in data science applications.

We write this document primarily for three audiences:

- For **data science researchers**, we hope to offer a roadmap, laying out the landscape of responsible data science research and highlighting important research questions that have not received much attention so far.
- For **funding agencies**, we aim to show where research is needed to advance the area of responsible use of data and algorithms and highlight the risks of leaving these subjects under-researched. We also highlight some of the current impediments in the larger academic culture to effectively perform research in this area.

- For **stakeholders** such as practitioners, end-users and policymakers, we want to highlight some of the most important things to be aware of when applying data science methods and some of the ways they can and should be involved with all phases of the data science process, including the academic research and the shaping of national and international policy.

In the first section, we consider several case studies highlighting different aspects of the irresponsible use of data and algorithms. In the second section, we use the case studies as examples for shaping a set of requirements that need to be addressed to make responsible use of data and algorithms a responsible practice. In the third section, we attempt to distil the most important general research challenges as we see them. Finally, we discuss what impediments we see to effective investigations into these issues, both practical and social, in the current research landscape. We make a series of suggestions for what changes can be made to better facilitate research into responsible use of data and algorithms.

1) Case studies

Artificial intelligence, machine learning, data science, data analytics, decision-support systems, among others, are becoming more and more predominant in our daily life. Such algorithms and models are often used to help us decide which items to buy, what music to listen to, but also in high-stakes domains such as education, healthcare provision, job recruitment, criminal justice, among others. However, there are also cases where they can create inequality and inequity, perpetuate bias and prejudice, and produce unlawful or unethical outcomes. Following, we review several case studies that highlight diverse aspects of irresponsibility regarding the use of data, algorithms or ethical and privacy concerns. It is important to note, however, that these case studies are not exhaustive or complete. While some of the issues emerging from the following case studies have been addressed or remedied, they can still affect or harm the actors involved.

Throughout history, the technology proved to be a powerful tool. Technology allows for the fast processing of large amounts of data, drawing inferences that might not be easily captured by humans, and usage at scale. However, irresponsibly using it might lead to undesired outcomes. The punch cards created by IBM to identify, organize and number the population are one such example from more than 75 years ago. It is believed, however, that these punch cards were used by the Nazi regime to identify Jews and Gypsies during the Holocaust. Similarly, around 1994, the nationality of the population in Rwanda was collected in so-called identity cards. These identity cards then prevented movements between the groups, and it was further believed that they were used in the Rwandan genocide to identify people with Tutsi ethnicity.

Data is always at the core of algorithms. However, historical data is known to encapsulate human and structural biases, implicit bias, prejudice, or be unbalanced with regard to the overall population. Thus, the population affected by such cases is more prone to be affected by incorrect predictions. We can name several case studies in which historical human biases led to irresponsible outcomes. We start with an example of a machine learning model which aims to identify patients who are most likely to be early discharged from the hospital (Rajkomar et al., 2018). The main finding of the data analytics group that developed the model was that adding the ZIP code of the patient improved the accuracy of the model.

However, they also found that when the patient lived in a socio-economically depressed or predominantly African American neighbourhood, the machine learning model predicted longer stays in the hospital. So, the algorithm would advise the hospital to provide additional care to a predominantly white population instead of a socially at-risk population. In the recruitment domain, Amazon developed a machine learning algorithm to predict which candidates' CVs should be considered as good candidates to hire.[1] However, the overrepresentation of male employees' CVs in the training data led the algorithm to associate male candidates as better options only because they were overrepresented in the training examples.

Besides historical biases, datasets used in the AI and data science landscapes also deal with the under-representation of minority groups and the lack of critical reflection about the capabilities and limitations of the datasets. For example, in 2015, the programmer Jacky Alciné discovered that Google Photos had automatically tagged a picture of him and a friend as "gorillas".[2] Alciné and others attributed this algorithmic output to a lack of diversity in the data on which the tagging system was trained. The tool LYNA — short for Lymph Node Assistant — was developed by Google to identify breast cancer tumors that metastasize to nearby lymph nodes (Steiner et al., 2018). That can be difficult for the human eye to see, especially when the new cancer growth is small. Being able to identify such nodal metastasis is imperative, as it impacts the course of action in terms of treatment, especially in breast cancer. In one trial, LYNA accurately identified metastatic cancer 99% of the time using its machine-learning algorithm. However, more testing is required before doctors can use it in hospitals, especially because there is still little information regarding the training/testing data and the number of patients that participated in the study, among others.

One other issue closely related to the datasets used in AI systems and algorithms refers to the misuse of sensitive or protected attributes for identification. Commonly, gender and race are among the most representative. It is clear, however, that eliminating them is not enough and may even have very harmful consequences. While Amazon did not use the gender of the applicants as a feature in their algorithm to predict which candidates' CVs should be considered for hiring, the algorithm was still able to discriminate based on hobbies and the use of gender-specific words. For example, for female candidates, the term "women'' is often found in the name of the clubs and educational institutes they joined or attended. Furthermore, male candidates are more prone to describe themselves with words such as "executed" or "captured". Similarly, the race was not given in the COMPAS algorithm, which is used by U.S. courts to assess the likelihood of a defendant to become a recidivist. ProPublica investigated the COMPAS tool, however, and found out that black people are almost twice as likely to be labeled as high risk compared to white people, but in fact, they do not appear to reoffend.[3] While neither of these case studies used information regarding race or gender, such inferences can still be drawn from proxy variables (postcode, income, hobbies), which strongly correlate with race or gender.

Proxy variables can affect some populations even in the absence of a machine learning algorithm by simply using information stored in cookies. This is likely happening due to the widespread digital footprints collected from users, customers, and in general, citizens. Such digital footprints, however, raise concerns about privacy and data ownership. Back in 2012, the online office retailer Staples was found to

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[2] https://www.wsj.com/articles/BL-DGB-42522
[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

vary product prices based on customers' ZIP code: lower product prices were given to customers whose ZIP code was less than 20 miles from a competitor's physical store.[4] It also seemed that the ZIP codes that received lower prices were usually located in wealthier neighborhoods. Similarly, in 2015 it was found that The Princeton Review was charging customers a different amount of money for their online tutoring, based on their ZIP code location.[5] Moreover, an investigation conducted by ProPublica also demonstrated racial bias --- The Princeton Review was charging Asian families higher prices than non-Asians families.

As we have seen so far, some case studies do not need sophisticated automation to make irresponsible use of data and algorithms. Still, they should make people reflect on the use of technology and the implications of using it. Even though there are cases in which our digital traces can be used for the greater good, such as identifying women with postpartum depression, people with high-risk depression, or with suicidal patterns, these inference algorithms deal with and emphasize privacy and data protection, confidentiality, algorithmic fairness and transparency, as well as ethical issues. To name some examples, even a little personal information, such as the place where the person lives, gender and date of birth, was easily used to identify 53% of the U.S. population (Sweeney, 2000). When looking at social networks, Facebook Likes were enough to predict with high accuracy various personal attributes such as race, gender, sexual orientation, political values, religion, use of substances and even whether the parents of an individual separated before their 21st birthday. The HireVue company, which analyzes candidates' CVs and identifies the best candidates, used to offer one product to their client companies: record the interviews and analyze the emotions and cognitive states of the candidates and then send this data to the employers.[6] The emotion analysis software used in this case is Affectiva, a software (and company) that focuses on understanding human emotions, cognitive states by analyzing their facial and vocal expressions. While Affectiva started as a product aiming to support children with autism, it was later used by the HireVue company.

Typically, algorithms are evaluated with various performance metrics, such as precision, recall, accuracy, and even engagement metrics. The choice of these metrics, however, can have a substantial impact on the fairness of the algorithm. In the past years, YouTube worked on solutions to mitigate the effects of recommending quantity over quality and, thus, promote quality over quantity.[7] One of their solutions was to recommend videos based on newly implemented "responsibility" metrics, which are meant to identify content that can not only go viral but is also constructive, using various human-in-the-loop approaches (content moderators, surveys). These metrics, however, are opaque to the public. Recommendations are also typically catered towards known preferences of users, so they might be biased in nature. Sure recommendations would emphasize and confirm people's biases. The two issues above mentioned, however, do not exclude each other.

Many case studies raise concerns about their overall design, choice of technologies, as well as the integration of societal, legal and ethical needs and concerns, especially when applied in high-stakes domains such as criminal justice or education. In 2020, for instance, a machine learning algorithm was developed in the UK to predict students' grades in the GCSE (General Certificate of Secondary Education)

[4] https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review
[5] https://techscience.org/a/2015090102/
[6] https://www.technologyreview.com/2020/02/14/844765/ai-emotion-recognition-affective-computing-hirevue-regulation-ethics/
[7] https://www.bloomberg.com/news/articles/2019-04-11/to-answer-critics-youtube-tries-a-new-metric-responsibility

exams to eliminate the concern that bias might arise if grades were predicted by teachers alone.[8] A statistical model was used to predict grades on a similar distribution as had been observed in previous years, rather than using the individual performance of the students. Features used for this prediction included historical averages of the school, which led, however, to inflated grades for prestigious schools. In consequence, high-performing students at underperforming schools were given lower grades, while average students at better schools were treated more leniently. In the end, the government decided to ignore the grades predicted by the algorithm.

Finally, many well-known case studies with irresponsible outcomes lack human oversight in several stages of the pipeline, as well as in the development phase. For instance, in the example of Google Photos, the irresponsible algorithmic output was also attributed to a lack of diversity in the programming and testing teams that produced the model. The Ministry of Social Affairs developed the Dutch System for Risk Indication (SyRI) to detect welfare fraud.[9] SyRI used an algorithm that combined and analyzed citizen data from several governmental agencies to identify possible fraud. Based on several risk indicators learned from previous fraudsters, the algorithm created a risk profile for every citizen. In case of an increased risk of irregularities, the algorithm marked the citizen as a possible fraudster. The cases identified by the algorithm were then checked by a human representative. However, we argue that human oversight should be an integral part of the system instead of a final verification step. While in the Netherlands, the use of SyRI was interrupted in 2020, such algorithms are widely used in the UK and the US, although it is known that they cannot distinguish between fraud and innocent mistakes.

2) What is needed for responsible use of data and algorithms?

We further use the case studies described in the previous section to identify some of the central research areas that they impact. Based on the technical, legal, ethical, and societal issues they draw attention to, we highlight several aspects that are needed to help shape the practice of using data and algorithms responsibly.

We need high-quality descriptions of data and datasets

Data and datasets need to be critically understood and characterized before being put into practice. While we acknowledge that determining the representativeness of data is not a trivial problem, we argue that data and algorithms practitioners should have a critical view on the datasets they use and analyze whether the training data used for their algorithms is representative and does not contain selection or collection bias, among others. Efforts to help in describing datasets and models in a more transparent way are currently receiving much attention, see for example, “Datasheets for Datasets”, ‘’Data Statements’’, among others. However, we argue that we need adequate, standardized methods to analyze and understand the distribution of features our datasets encapsulate.

It is also important to understand to what extent such datasets reflect real-world populations (see, for

[8] https://en.wikipedia.org/wiki/Ofqual_exam_results_algorithm
[9] https://www.privacyfirst.eu/court-cases/tag/System%20Risk%20Indication.html

example, the Amazon recruitment case study). For each created dataset, researchers and practitioners need to clarify how the dataset was collected, what is included and excluded from the dataset, and provide any other details regarding how and by whom the labelling of the data was performed. Having these detailed descriptions would allow to address in a pragmatic way the issues of data coverage and representation. Thus, detailed descriptions of datasets must become a requirement in light of increasing responsibility and shading light on the capabilities, limitations, and possible bias, inequality, and inequity of the solution (see the case study of LYNA, which provides limited information with regard to data).

In addition, we argue that researchers and practitioners need to document how the data have been used and critically assess how suitable they are for the use case at hand. A suitable example for this is the case study of General Certificate of Secondary Education, where students' grades were predicted based on historical school averages instead of their performance. While datasets could potentially contain rich features and attributes, it is also necessary to understand and identify the accurate and relevant factors/features that a machine learning algorithm can rely on.

Research has been focused on de-biasing techniques for training data. It is also a known fact that humans are biased and human-annotated and human-generated datasets are biased, see, for example, the case study regarding patients discharged from the hospital. Thus, it is important to understand which are the protected groups or sensitive attributes, as well as potential correlations in the training datasets for machine learning models, even before deploying the model.

We need safe, robust, and reliable methods to perform responsible data science

Algorithms need to be evaluated beyond accuracy metrics to encapsulate ethical, legal and societal values, among others. The mainstream approach for evaluating the success of an algorithm or a model is to measure its performance in terms of precision, recall or accuracy, against a benchmark. In high-stakes domains, we see, however, that such methods are prone to perpetuate bias, prejudice, inequality, and inequality. Thus, the success metric that we choose might have a substantial impact on the fairness of the algorithm. At times, the use of evaluation metrics beyond the standard might involve the need to develop new evaluation metrics.

Designing and applying proper methodologies for evaluating AI systems also allow us to reason whether the system performs equally well or has similar outcomes on different groups of individuals with different characteristics. Besides performance metrics, suitable tests, studies, or metrics should be developed in order to evaluate and measure the equity, equality or fairness between and across groups and individuals. One main concern that needs to be tackled is to what extent outliers in our data are handled by the methods that we implement. Thus, we argue that systematic evaluations of randomly selected outliers in the distribution could help us understand how algorithms fail to generalize (see Google Photos case study). Similarly, we could learn about the misleading variables that were used in decision making.

A critical reflection on the success metrics is also needed in the current digital society, in which citizens constantly turn to acquiring and contextualizing information from online sources. In this regard, decision-support systems play a pivotal role. On the one hand, recommender systems should provide

recommendations that are not solely content-oriented or that create inequity among content providers or content creators. At the same time, efforts should focus on understanding how the user consumption of a recommender system output changes with, for example, more content diverse recommendations, more topic diverse recommendations, among others, and study to what extent users' behaviour is modifiable. On the other hand, quality and quantity metrics should be fairly balanced. This would ensure that low-quality articles are not massively spread as a result of optimizing a recommender platform solely on engagement.

Sometimes, the ideal data do not exist and yet are sufficient for methodological research purposes. However, the accuracy of the method in the target population should be examined through coverage criteria and should be evaluated with counterfactual reasoning against standardized adversarial attacks, including adversarial learning. In terms of decision-making algorithms, we again argue for their robustness and reliability. Such algorithms should have the ability to not provide an answer when the predictions are unreliable, unsure or have low confidence and provide an explanation when they do provide a prediction.

We need transparent, reproducible, and generalized research results

Efforts to make data and models FAIR, to help describe them in a more transparent way, are gaining tremendous popularity, see, for example, "Datasheets for Datasets" and "Google Cloud Model Cards". The Google Cloud Model Cards should be used as a first step to improve the transparency of machine learning models, which is closely related to making models FAIR. We do argue, however, that such models could be further improved. For instance, the model cards provide no characterization of where the algorithm consistently fails to perform its task.

We, as researchers and practitioners, need to contribute to open science by building open source software and offering the community at large the opportunity to test, adapt or extend existing frameworks. Similarly, we need to make use of independent evaluation frameworks such as OpenML, which provides a collaborative environment to share, evaluate, reuse, and further analyze datasets and machine learning models.

As users, we need control over our data

The widespread digital footprints that are collected from users, customers and, in general, citizens, raise concerns with regard to privacy and data ownership. While people may not choose to publicly share personal information, the extensive records of online behaviour can be still used for prediction, leading to privacy intrusion and the public disclosure of private facts (Kosinski et al., 2013). More precisely, in the Staples example, customers were not required to add their ZIP code, but their location was inferred through their IP address (stored in cookies).

The General Data Protection Regulation (GDPR) provides a framework for the collection and processing of personal information for those individuals that reside in the European Union. The main aspects covered by GDPR are for businesses, companies, websites, etc. to: (1) provide transparency with regard to data use, (2) provide data subjects with access to personal information collected, (3) allow requests to delete or correct personal information and (4) allow people to restrict the access to data processing.

GDPR managed to put the focus on the rights and responsibilities of data subjects and processors of personal data. It is important, however, to make such responsibilities transparent, in a manner that is accessible and understandable for the large population. While explainability is often understood as a manner of informing citizens on how inputs and outputs of a system are related, this may not necessarily also increase citizen control over their data and how their data is used by various algorithms (Edwards and Veale, 2018). Thus, it is still vital to make citizens aware of the information they provide when interacting with online services and automated decision-making.

Critical reflection on the task in practice and as practiced by people vs machines

All stakeholders involved in the design, implementation and application of AI systems should critically reflect on how to operationalize data protection, confidentiality and transparency in given settings. Furthermore, there is a need to properly evaluate the risks of making inferences or predictions that can affect the privacy of individuals. We argue that the element of responsibility becomes more important when transitioning from research to application. Authors should carefully delimit their claims, and editors and reviewers should hammer down on weak or unsubstantiated claims of their fitness in real-life applications.

It is known that the algorithms we create are prone to reflect, at times, even our own (hidden) biases. Thus, people should be aware of their own biases in order to be able to build responsible systems. For responsible use of data and algorithms, AI practitioners and researchers should be composed of mixed teams, while representative stakeholders should be involved at any stage in the lifecycle of an algorithm, prototype, or product. Consulting with stakeholders may help to identify and eliminate undesirable algorithmic behaviors and help to correctly position the algorithm in its real-world context. Looking at the examples of the Dutch System for Risk Indication and the Correctional Offender Management Profiling for Alternative Sanctions, bringing humans-in-the-loop could have mitigated, eliminated or early identified the problems that might arise from connecting several datasets with potentially inaccurate or missing information. Furthermore, human oversight is also necessary in order to understand which are the risks and impacts associated with all the stakeholders involved. Such threads can, thus, be identified before deployment in an open environment.

Responsible AI systems should be designed with the goal of taking human values into account, such as fairness, accuracy, confidentiality and transparency, in order to contribute to a better society. However, even before the design phase, we need to reason on the suitability of using, for example, predictive models in high-stakes domains, such as the case of the General Certificate of Secondary Education, in the UK. Northpointe mentions that the COMPAS system is racially neutral and that the rate of accuracy is similar for both black and white defendants. Thus, they argue that the outcome is not biased because the amount of correct predictions is similar for both groups. Several studies, however, concluded that the fairness metrics implemented in the COMPAS tool are incompatible, and the outcome of the tool is nevertheless prone to bias and discrimination. This leads to another requirement for designing and developing AI systems, namely ensuring that anti-discrimination laws are not violated, and that cultural, ethical, social-economic biases or stereotypes are neither perpetuated nor emphasized. Nevertheless, one interesting aspect of the development of algorithms that use historical data is that they reveal the

inherited bias and discrimination in human decision making. Thus, we argue that at least such outcomes should serve as a starting point to reexamine how people make decisions in the first place.

The research and business landscapes are constantly and rapidly evolving, and the abundance of available data encourages research in various domains, among which computer vision. Besides the emotion analysis case study (Affectiva and HireVue), we observe an increasing number of technologies that deal with criminality prediction or terrorism (the case of the startup Faception), homosexuality prediction (Wang and Kosinski, 2017), among others. Such examples, however, emphasize the need to understand what kind of criteria, norms and legal aspects should be met by inference-makers to make predictions in the boundaries of ethics. Furthermore, the AI landscape should consider whether these criteria, norms and legal aspects are or should be different for research, education or commercial purposes. Nevertheless, the potential impacts that might result from deploying or simply developing a particular technology should be explored in a systematic way and become a requirement for responsible use of data. Guidelines should be created in order to critically reflect on the consequences of performing research in sensitive areas, as well as further using such research in commercial or governmental institutions (see the case study of Affectiva and HireVue). Such guidelines should be equally important for both scientific and practical use of data science. On the one hand, in the scientific domain, we argue that together with the open science movement, which strives for making data and code publicly available and easily accessible, it is the responsibility of technology developers, disregarding their field and background, to provide social or ethical impact statements. On the other hand, in the industry or practical use of data science, we argue that besides social and ethical responsibilities, the responsibilities of providing qualitative and transparent solutions should also be met. In the case of the latter, the role of the scientific community would be to provide clear guidelines of how these responsibilities should be tackled.

3) General challenges

In the previous sections, we highlighted a number of real-world use cases and identified a number of requirements for fostering, advancing and establishing the use of data and algorithms in a responsible manner.

However, for a research field to truly gain momentum, it is important also to have larger challenges - those questions that take a generation or more to answer, or perhaps, those that may never be solved.

C1: Standardized evaluative criteria for responsible data science

In the past couple of years, the computer science literature saw a growing number of papers introducing and discussing various definitions of fairness. For a review, see (Caton & Haas, 2020), (Mehrabi et al., 2019). This, however, is a natural consequence of the fact that for fostering and performing responsible use of data and algorithms, one needs to consider multiple factors, such as the context of the problem, the domain application, as well as the needs of many stakeholders. In his tutorial on “21 definitions of fairness and their politics”, Arvind Narayanan states that a single fairness metric is not sufficient to build algorithms that facilitate and encapsulate human values.

In (Fazelpour & Lipton, 2020), the authors outline an important distinction between different approaches to fairness in machine learning. *Ideal* approaches are those that state, up-front, what constitutes a fair mechanism, and then proceed to correct an unfair mechanism to make it fair. This tends to lead to a kind of closed-world approach where the problem is treated in isolation, and anything that improves the rigid definition of fairness is seen as beneficial. *Non-ideal* approaches take a larger view of the world and acknowledge that any intervention to improve fairness has other consequences, that intervening without understanding the underlying causal structure of fairness may have unintended consequences, and that actions cannot be made fair without studying the dynamics of the larger system.

While such a holistic approach is more honest about the complexities of the problem, it also leads to a conception of the world that is far more difficult to formalize. This is not a problem in general for political science, since there is usually a less strict requirement for formality in that domain. However, for interventions to be implementable in computers, a much more formal treatment is required. Thus, one of the challenges that we envision is how do we build generally agreed-upon frameworks for representation and computation, not just for fair ML, but for all responsible data science, in such a way that they can be implemented in autonomous or semi-autonomous systems.

C2: Conventions for responsible research, publication and production.

As the case studies section shows, there is no shortage of examples of irresponsible use of data and algorithms. However, in almost all cases, the practitioners and researchers involved appear to have acted in good faith. Assuming that a practitioner or researcher is in principle willing to spend the time and effort required to make their use of data and algorithms responsible, what tools can we offer them to help them avoid irresponsible outcomes?

One important aspect of this problem is informing external stakeholders of risks. For instance, it may be very helpful to inform policymakers of the potential long-term impacts of a particular piece of research when it is published so that a slow-moving legislature is given the maximum amount of time to respond to potentially disruptive innovations.

Initial steps have been made in this area. For the publication of models, the practice of model cards (Mitchell et al., 2019) has been proposed. For the publication of data, various approaches have been proposed (Bender & Friedman, 2018), (Gebru et al., 2018), but they are still not standardized. The recent movement towards encouraging or requiring broader impact statements in AI research may well make the literature easier to follow for policymakers. However, with all these, there is limited knowledge about how exactly these solutions should be best implemented. Contrast this with the writing of related work or method sections, where strong conventions exist.

Finally, given the pressures of publication and deployment of production systems, it may be desirable to offer tools that can be used earlier, both in the research and practical (i.e., industry) pipelines, when it is still easy to pivot towards a more responsible approach. Simple checklists or questionnaires may offer a structured way to think about responsibility at stages like tendering for a data science project, evaluating a research proposal, or during a researcher's or practitioner’s yearly evaluation.

C3: Structuring research beyond publication.
In most areas of data science research, a publication is what counts as the final result of a research project. A system is built, evaluated, presented, and then the researchers move on to follow-up research or other topics. In other domains, there are often larger trajectories consisting of multiple publications. In many life sciences, a statement is accepted as scientific truth not when a single experiment confirms it, but when a meta-analysis of a large body of experimental work confirms it. In medicine, drugs move through several stages of trials before they are accepted as safe and effective. In social psychology, both pre-registration and independent replication are important steps.

If we compare this with a data science discipline like machine learning, we see that there is far less structure beyond a series of publications, which often consist purely of presenting a model that achieves higher performance than some other model. Recently, many papers have shown that published increases in performance were either illusory or not caused by the mechanism suggested in the paper [Lucic 2018, Melis208, Ruffinelli 2020].

For the subfields of data science to retain their flexibility but also achieve greater maturity, it may be necessary to keep the requirements for publication as they are but reduce our trust in those assertions that have only been studied in a single publication. In this way, we can add additional levels of confidence for work that has not only passed peer review but also survived meta-analysis or has been independently replicated. Exactly what form these additional validations should take is likely highly dependent on the specific domain and should be thoroughly studied.

C4: Balance performance and responsibility.
Human values, such as fairness, accuracy, confidentiality, and transparency, should be at the core of developing and using algorithms in a responsible manner and thus contributing to a better society. Current publication practices as well as evaluation practices in various computer science subfields, however, are still significantly driven by performance metrics. As we saw in our case studies, in high-stakes domains, precision, recall, or accuracy metrics are prone to perpetuate bias, prejudice, inequality and inequality. Nevertheless, simply considering performance metrics can not ensure that anti-discrimination laws are not violated, and that cultural, ethical, social-economic biases or stereotypes are neither perpetuated nor emphasized. Thus, balancing performance and responsibility is a challenge to be addressed in the landscape of responsible use of data and algorithms.

4) Landscape of Responsibilities

In this section, we cover some of the challenges we see in the wider academic landscape to make responsible use of data and algorithms a reality. We start by looking inwards, at what we as academics may need to change in our outlook and expectations and work out step by step to cover the challenges in research collaborations, university departments and finally, national and international entities such as funding agencies and policymakers.

The computer science outlook

Data science is a diverse mixture of disciplines. Given that the authors of this article come primarily from a computer science background, that is where we start.

As computer scientists, we are used to a particular set of assumptions that may not always hold when studying the subject of responsible data science or responsible use of data and algorithms. For instance, we commonly abstract problems to a general formulation, which is then (hopefully) solved and applied to each specific instance of the abstract problem. In machine learning, for instance, the tasks of character recognition, evaluating chess problems and recognizing traffic signs can all be abstracted to the problem of *classification*. For this problem, several algorithms exist, which can then be applied to each.

Another example is the storage of relational data. It does not matter what we are storing, be it customer data, website content or experimental results. So long as we can find a suitable mapping to the basic structure of a relational database, we can solve the problems of storage and query on an abstract level. This kind of abstraction and separation of concerns is behind many of the greatest accomplishments of computer science.

There is no guarantee that such approaches based on abstractions will work equally well where social and human concerns are involved. Abstracting problems may lead to quantification fallacies and de-emphasize poorly measurable human impact. Indeed, there are many historical examples available [O'Mahony 2017] of over-automation in business and government, where the assumptions of system engineers proved flawed, and those people who formed exceptions to the rule suffered as a result. Responsible data science is asking us to consider the potential application of the method, and therefore, a responsible data scientist will critically examine how the model fails to produce the desired result. Such aspect should no longer be neglected or de-emphasized.

One source of inspiration may be the discipline of human-computer interaction (HCI), which is equally relevant as a scientific discipline and as an engineering practice in the design of systems. In the scientific environment, while this is a sub-discipline of computer science, it commonly deals much less in rigid abstraction. Interfaces are always purpose-built and commonly evaluated with all stakeholders involved in the process. The emphasis in the world of interface design is often on non-quantitative, inclusive processes like co-design, user-centred design and lateral thinking.[10]

Outside of academia, HCI solutions can be found inside companies developing applications and in specialized design firms. What is usually abstracted across projects is not a set of prebuilt solutions but a set of processes, methods and design patterns: generally applicable principles and solutions that can be flexibly applied when appropriate. It is only at this level of abstraction that academia comes in: codifying and evaluating design processes and design patterns.

The social science outlook

Most of the problems around responsible data science and AI result from trying to understand or learn human behaviour or its trends by using sensitive, personal or historical data. As seen in our previous case studies, systems that are trained with historical data or human- or society-gathered data inherit and then automate the bias found in our society. Thus, even when faced with a well-thought algorithm trained

[10] Quantitative methods like A/B testing are also common in design, but there is generally a strong awareness of the pitfalls of the quantitative fallacy [Chen, Domingues 2020].

with a perfectly collected dataset, we may still act unfairly to certain social groups simply because society treats them differently or in an unfair way. This suggests that there is an obvious benefit in looking to the social sciences for guidance and inspiration to assure that potential biases are not encoded in our data and algorithms.

The social sciences are more accustomed to working in a bottom-up fashion, without expectations of rigid generalization. In the domain of AI safety, multiple calls exist for more inclusion of the social sciences in developing AI that aligns with human values [Irving 2019].

The ethical, legal and societal outlook

For society to benefit from technological and research advancements, we need to be able to communicate with policymakers and help them translate the research we perform into policies. Thus, it is necessary to give policymakers the necessary tools to anticipate and understand existing or future research. This would allow them to implement effective laws and other measures to regulate new technology, make effective use of technology as well as containing its applications. There are, however, certain criteria to acknowledge and be dealt with in order to facilitate the communication from researchers to policymakers, and similarly, from research to policymaking. First of all, our research community is mostly focusing on developing novel and original contributions. Policies, however, are typically not a one-time thing, but they should be generalizable and applicable across a broader range of problems. Second, most computer science researchers work under a certain subfield, with little or no multidisciplinary interactions. Policymaking, on the other hand, requires multidisciplinary perspectives.

In the academic or research landscape, broader impact statements have recently gained some momentum. For instance, 2020 researchers were obliged to write about the positive and negative impact of the research they aim to publish when submitting to the NeurIPS conference. One consequence of this scheme is that policymakers could look at the collection of these statements in aggregate to keep abreast of technological development and its potential impact without needing to study the scientific literature in detail. Moreover, we argue the necessity of introducing and applying technology maturation levels. Researchers should make use of grants to move research into more mature phases, which include greater stakeholder inclusions, more robust solutions, among others.

There are many pitfalls to this scheme. There is currently no evidence that broader impact statements are followed by policymakers or similar stakeholders. There is little consensus on how these statements should be written, and their effectiveness relies purely on the ability of the researcher to guess at the impact of their work. Nevertheless, it shows one small step towards allowing a greater degree of communication between the developers of new technologies and those impacted by them.

The area of legal scholarship perhaps offers a unique perspective, which may help to bridge the outlooks of the AI and computer science communities. It shares with (other) social sciences the broad remit, deeply tied to social structure and human behavior. With computer science, it shares the need for strict definitions and rules that can be implemented by a system that, while not autonomous, operates in principle in a way that is divorced from personal values and perspectives. Reviewing legal scholarship in this right may offer valuable insight into how we can codify human values into a rigid enough form that it can guide the responsible use of data and algorithms, and yet remain flexible enough that it does not become a value in itself, rather than a distillation of human values.

The interdisciplinary outlook

It is widely acknowledged that interdisciplinary research is rife with challenges. These start simply in getting researchers with different backgrounds to collaborate. Each speaks their own language and has their own worldview, and integrating these is hard work. Often, this work is left to the junior scientists, who are tasked with performing the interdisciplinary research under the supervision of senior researchers from the fields to be integrated. In such cases, it falls to the junior researchers to shape the language and view of their senior peers. A dynamic that usually flows in the other direction.

If the researchers do manage to build a common vocabulary and produce research, they will likely run into the problem that different fields have different criteria of success. What in one field counts as a valuable artifact, may be eyed with little more than suspicion in others. This is not just a problem for the researchers in deciding a common goal to work towards. It also throws up boundaries for getting work accepted. If reviewers are not accustomed to the styles of evaluation required by interdisciplinary research, the already low probability of getting work accepted may drop considerably compared to the colleagues who were wise enough not to engage in interdisciplinary research.

Even if the work gets accepted, university departments and hiring committees may not know how to value research from other fields. For instance, the conference publications that make up the bulk of computer science research may be dismissed in other fields as simple talks, nowhere near as important as journal publications.

The conclusion is that when it comes to interdisciplinary research, the deck is stacked against us. There are many calls from universities and funding agencies alike for interdisciplinarity, but little acknowledgement of the poison pill it can represent, especially for young researchers. To truly stimulate interdisciplinarity, the odds should be evened, and frameworks must be built where somehow, interdisciplinary research is given a greater benefit of the doubt without reducing the rigour of academic scrutiny that it is subjected to.

While we do not aim to solve the problem of interdisciplinary research, we consider the following requirements especially pressing in the area of responsible data science.:

- A clear framework of what constitutes valuable interdisciplinary research.
- Publishing venues such as workshops, conferences and journals which honors these values
- Career evaluation that honors these values.

From a computer science perspective, a legal analysis or a quantitative study will likely seem un-empirical or insufficiently rigorous. For interdisciplinary research, it is especially important that those performing the evaluations are sufficiently familiar with the methodology being used to evaluate the work in question.

The institutional outlook

One of the many problems highlighted by the childcare benefits scandal is the manner in which institutional culture can feed into the problems of irresponsible data science. Many of the problems that led to the benefits scandal were entirely non-technological in nature: institutional discrimination, lack of oversight, and a one-sided focus on catching fraud. However, within such a culture, there are no safeguards against the irresponsible use of data science methods. Any method that promises results in

the single metric of the amount of fraud detected is likely to be implemented without careful attention paid to algorithmic transparency and inadvertent side effects. In this context, however, the institution itself should not be the only one held accountable. Institutional responsibility should also come from inside, from the humans conducting the duties of the institutions.

One route to more careful use of data science is to create stronger ties between practitioners at institutions such as businesses and governmental organizations and researchers in academia. Such ties are already very strong for data science research in general, but they are most often focused on using data science to optimize various quantitative metrics like profit, efficiency and retention.

For more qualitative aspects like responsibility and culture, the connections are much less strong and building such connections is likely more difficult. Allowing academics to sit in judgment over your business processes is a hard sell. To make this a proper, two-sided conversation that benefits both parties, we can set the following goals:

- Practitioners should help academics set relevant research questions by helping them to understand the practical implementation choices they face.
- Academics should help practitioners to detail the potential effects of their deployments. This can be done through case studies of similar systems deployed in the past, small scale simulations, or additional metrics such as bias measurements. Both academics and practitioners should work together to make these as simple and compelling as possible so that they can be used to influence decision making beyond just the data science department.
- Academics should make the language of organizational decision making their own and frame the risks of irresponsible data science in those terms. It's a common fallacy to think that business decision making is purely motivated by profit, or that qualitative metrics are always ignored. Most institutions are careful about managing long term risks and working to optimize qualitative metrics like employee and user happiness. If the tradeoffs of responsible and irresponsible data science are framed in these terms, rather than as a moral imperative, not only are we more likely to see an uptake in organizations, we are more likely to see a careful and insightful balancing of different requirements.
- To some extent, institutions need to be held to account. This needs to be done in a way that is effective but also fair. Rules and audits should not, for instance, disadvantage smaller companies more than larger companies, who have the capacity to deal with the overhead. A fair system can be devised only by a careful conversation between practitioners, policymakers and academics. Setting the rules and the method by which they are enforced requires a very deep insight into the realities of applying data science, the technological details of currently available methods and the realities of setting policy. Only by bringing together three types of professionals in an honest conversation can we hope to develop a fair and effective system.
- In standard interaction design practice, there is a strong set of principles for how to involve users in the design process. Through tools like user testing and cooperative design, institutions can offer their users a seat at the table and work together on designing the kind of product that the user would want to use. When it comes to matters of responsibility, however, the aims of the institution and those of the user may be more at odds. For instance, a social media company may want to maximize its access to a user's private data, while a user may want to control how much such data is made available. Here, a disinterested party like academia may be helpful by

studying such processes and distilling successful patterns and design methods from both perspectives.

Summary

This document presents a roadmap for research on responsible use of data and algorithms, driven by several case studies that highlight different aspects of irresponsible use of data and algorithms and discuss their biased, unethical and unlawful outcomes. We used these case studies to identify a set of requirements that we believe need to be addressed by both data science researchers and practitioners to make responsible use of data and algorithms a responsible practice. Challenges, however, come with requirements and responsibility. Thus, we also make an attempt to extract some general research challenges that we consider to be important. Finally, we make a series of suggestions for what changes can be made to better facilitate research into responsible use of data and algorithms, which cover several landscapes such as computer science, social science, ethical, legal and societal, among others.

References

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Steiner, D. F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., ... & Stumpe, M. C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology*, *42*(12), 1636.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *Proceedings of the* 5 th *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Fazelpour, S., & Lipton, Z. C. (2020, February). Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 57-63).

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring fairness in machine learning to advance health equity." *Annals of internal medicine*169, no. 12 (2018): 866-872.

Sweeney, Latanya. "Simple demographics often identify people uniquely." *Health (San Francisco)* 671, no. 2000 (2000): 1-34.

Kosinski, Michal, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." *Proceedings of the national academy of sciences* 110, no. 15 (2013): 5802-5805.

De Choudhury, Munmun, Scott Counts, Eric J. Horvitz, and Aaron Hoff. "Characterizing and predicting postpartum depression from shared facebook data." In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 626-638. 2014.

Irving, Geoffrey, and Amanda Askell. "AI safety needs social scientists." *Distill* 4.2 (2019): e14.

Thomas, Rachel, and David Uminsky. "The problem with metrics is a fundamental problem for AI." *arXiv preprint arXiv:2002.08512* (2020).

Caton, Simon, and Christian Haas. "Fairness in Machine Learning: A Survey." *arXiv preprint arXiv:2010.04053* (2020).

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

Narayanan, A. 21 Definitions of Fairness and Their Politics. In *Tutorial Presented at the First Conference on Fairness, Accountability, and Transparency (FAT*) 2019*.

De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag, Autoriteit Persoonsgegevens 2020 https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf

O'Mahony, S. (2017). Medicine and the McNamara fallacy. *The journal of the Royal College of Physicians of Edinburgh*, *47*(3), 281-287.

Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019) *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* London: Nuffield Foundation.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2017). Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*.

Melis, G., Dyer, C., & Blunsom, P. (2017). On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.

Ruffinelli, D., Broscheit, S., & Gemulla, R. (2019, September). You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*.

Chen, A. Does AB testing lead to crappy products?
https://andrewchen.com/does-ab-testing-lead-to-crappy-products/

Domingues B (20202, Feb)The problems of over-relying on A/B testing
https://uxdesign.cc/3-silent-problems-of-overlying-on-a-b-testing-203c3887f1c1

A1) Case Studies

1. *Discharging patients from hospitals*

Researchers in the data analytics group at the University of Chicago Medicine created an algorithm to identify patients who are most likely to be discharged early (Rajkomar et al., 2018). Such an algorithm would allow the hospital to provide additional case management resources to these patients, to ensure that they indeed would be discharged from the hospital sooner. The data analytics group developed the models using historical, clinical data. One of their main findings was that adding the ZIP code where the patient lives improved the accuracy of the model in identifying those patients who would have shorter lengths of stay. However, it was also found that when the patient lived in a socio-economically depressed or predominantly African American neighbourhood (based on the ZIP code), the machine learning model predicted longer stays in the hospital. So, the algorithm would have led to the paradoxical result of the hospital providing additional care to a predominantly white population, instead of to a more socially at-risk population.

2. *Amazon recruitment*

In the recruitment domain, Amazon developed a machine learning algorithm to predict which candidates' CVs should be considered as good candidates to hire.[11] However, the overrepresentation of male employees' CVs in the training data led the algorithm to associate male candidates as better options only because they were overrepresented in the training examples. While the gender of the applicants was not considered as a feature, the algorithm learned to discriminate based on hobbies and the use of several words that are more specific to a gender. For example, for female candidates , the term "women" is often found in the name of the clubs they used to join or in the name of the educational institutes they followed. Furthermore, male candidates are more prone to describe themselves with words such as "executed" or "captured".

3. *General Certificate of Secondary Education*

A more recent example, in 2020, a machine learning algorithm was developed in the UK, to predict students' grades in the GCSE (General Certificate of Secondary Education) exams, to eliminate the concern that bias might arise if grades would be predicted by teachers alone.[12] A statistical model was used to predict grades on a similar distribution as had been observed in previous years, rather than using the individual performance of the students. Features used for this prediction included historical averages of the school, which led, however, to inflated grades for prestigious schools. In consequence, the algorithm created inequality among students - the grade of the student was decided based on their

[11] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[12] https://en.wikipedia.org/wiki/Ofqual_exam_results_algorithm

postcode and the previous performance of the school, and not based on their performance in the classroom. Thus, high-performing students at underperforming schools were given lower grades, while average students at better schools were treated more leniently. Ultimately, expert advice was offered by the RSS (Royal Statistical Society), but withdrawn due to prohibitive NDAs. In the end, the government decided to ignore the grades predicted by the algorithm.

4. SyRI

Dutch Systeem Risico Indicatie (SyRI), System Risk Indication, was developed by the Ministry of Social Affairs to detect welfare fraud.[13] SyRI uses an algorithm that combines and analyzes citizen data from a large number of governmental agencies, to identify possible fraud. Based on certain risk indicators, learned from previous fraudsters, the algorithm created a risk profile for every citizen. In case of an increased risk of irregularities, the algorithm marked the citizen as a possible fraudster. The cases identified by the algorithm, were then checked by a human representative. UN-Rapporteur found that the system was prone to violate human rights because the system seemed to target especially poor and vulnerable groups of people, with a migration background, in Dutch society. While in the Netherlands the use of SyRI was interrupted in 2020, such algorithms are widely used in the UK and the US. It is known, however, that they cannot distinguish between fraud and innocent mistakes.

5. COMPAS

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool is a case management and decision support tool developed and owned by Northpointe (now Equivant).[14] COMPAS is used by U.S. courts to assess the likelihood of a defendant to become a recidivist. ProPublica investigated the AI tool and found out that black people are almost twice as likely to be labeled as high risk, compared to white people, but in fact, they do not appear to reoffend. Conversely, white people are more likely to be labeled as lower-risk than black people, but they do end up committing other crimes. ProPublica also found that only 20 percent of people predicted to commit violent crimes actually went on to do so. While the COMPAS software does not contain information regarding race, such inferences can still be drawn from proxy variables (such as postcode, income), which strongly correlate with race.

6. LYNA

Google developed a tool called LYNA — short for Lymph Node Assistant —, for identifying breast cancer tumors that metastasize to nearby lymph nodes. That can be difficult for the human eye to see, especially when the new cancer growth is small. Being able to identify such nodal metastasis is imperative, as it impacts the course of action in terms of treatment, especially in breast cancer. In one trial, LYNA accurately identified metastatic cancer 99 percent of the time using its machine-learning algorithm. This, however, does not mean that the algorithm is ready to replace pathologists. More testing is required before doctors can use it in hospitals, especially because there is still little information regarding the training/testing data, the number of patients that participated in the study, among others.

7. Google Photos

In 2015, the programmer Jacky Alciné (2015) discovered that Google Photos had automatically tagged a picture of him and a friend as containing “gorillas”. Alciné and his friend are African American, and their experience is one of the more straightforward and well-known examples of the racist outputs of

[13] https://www.privacyfirst.eu/court-cases/tag/System%20Risk%20Indication.html
[14] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

contemporary algorithmic systems. Alciné and others attributed this algorithmic output to a lack of diversity in the data on which the tagging system was trained and a similar lack of diversity in the programming and testing teams that produced it.

8. *Price discrimination*

Back in 2012 even, the online office retailer Staples was found to vary product prices based on the ZIP code of the customers. It was found that lower product prices were given to customers whose ZIP code was within approximately 20 miles from a competitor's physical store. In fact, it also seemed that the ZIP codes that received lower prices were located, in general, in wealthier neighborhoods. Similarly, in 2015 it was found that The Princeton Review was charging customers a different amount of money for their online tutoring, based on their ZIP code location. Moreover, the investigation conducted by ProPublica also demonstrated racial bias. It turns out that, in general, The Princeton Review was charging Asian families higher prices than non-Asians families.

9. *HireVue and Affectiva*

HireVue is a company that offers several products to companies, in terms of analyzing candidates' CVs and identifying the best candidates. HireVue claims that the analysis they perform allows companies to have a more diverse set of candidates, hire more diverse people, and thus eliminate potential bias. What is actually happening is that people are recorded during their interviews, likely without signing any consent form. HireVue analyzes their emotions and cognitive states during their interview, and then sends this data to the employers. The emotion analysis software used in this case is Affectiva, a software (and company) that focuses on understanding human emotions, cognitive states, by analyzing their facial and vocal expressions. While Affectiva started as a product aiming to support children with autism, it is now also used by the HireVue company.

10. *Recommender Systems*

Several social media platforms, such as Facebook, Twitter, YouTube, act as feed aggregators, recommending various types of items to users, such as news, videos, articles, among others. Such social media institutions are known to optimize their recommendation engine purely for one metric, such as engagement or number of views. This causes more radical content to be recommended more often. In the past years, YouTube worked on solutions to mitigate the effects of recommending quantity over quality and thus promote quality over quantity. One of their solutions was to recommend videos based on newly implemented "responsibility" metrics, which are meant to identify content that can not only go viral, but it is also constructive, using various human-in-the-loop approaches (content moderators, surveys). Such metrics, however, are opaque to the public and there are minimal details with regard to their implementation.

Recommendations are also typically catered towards known preferences of users, such systems might provide mainly biased items, thus emphasizing and confirming their biases. The two issues, however, do not exclude each other.

11. *Digital traces*

People leave numerous traces through their online behaviour and interaction with various social media platforms, websites or applications. Big data techniques can be used to harness these data and infer,

potentially with a high degree of confidence, information that can not be seen at first. These digital footprints can be used by algorithms to infer latent information about users such as pregnancy status, political affinity, sexual orientation, ethnicity, religious and political views, mental health conditions, among many other conditions.

Even a few personal information, such as the place where the person lives, gender, and date of birth, was easily used to identify 53% of the US population (Sweeney, 2000). When looking at social networks, Facebook Likes were enough to predict with high accuracy various personal attributes such as race, gender, sexual orientation, political values, religion, use of substances and even whether the parents of an individual separated before their 21st birthday. The retail network Target used the customers' shopping records to predict pregnancy and the due date in order to then send well-timed and targeted offers. Using a series of statistical data, (De Choudhury et al., 2014) leveraged social media posts to identify cases of postpartum depression of new mothers. Similarly, Facebook (de Andrade et al., 2018) has been working in the past years, together with domain experts, to prevent suicide and identify people that are in distress, based on their interaction with the platform, in particular messages posted online.