The AAAI 2017 Spring Symposium on
Artificial Intelligence for the Social Good
Technical Report SS-17-01

Identifying Sources of Discrimination Risk in the Life Cycle of Machine Intelligence Applications under New European Union Regulations

Syed Ali Asad Rizvi and **Elmarie Van Heerden** and **Arnold Salas** and **Favour Nyikosa**
Machine Learning Research Group, Oxford-Man Institute of Quantitative Finance, University of Oxford
{syed.rizvi,elmarie.vanheerden,arnold.salas,favour.nyikosa}@eng.ox.ac.uk

Stephen J. Roberts and **Michael A. Osborne**
Machine Learning Research Group, Oxford-Man Institute of Quantitative Finance, University of Oxford
{sjrob,mosb}@robots.ox.ac.uk

Elmer Rodriguez
ESADE Business & Law School, CaixaBank
elmer.rodriguezmendoza@gmail.com

Abstract

If machine intelligence systems acquire discriminating bias via data and underlying algorithms then this can introduce significant risks for members of the public subject to these systems. We outline two key sources of bias in machine intelligence applications that have become imminent in light of new General Data Protection Regulations introduced by the European Union in April 2016. These regulations introduce new and broad scope data protection laws regarding "any form of automated processing of personal data", which will directly impact the routine use of large collected public data sets and their use in automated decision making. This has direct consequences for introducing data bias if decision makers circumvent this law by avoiding data collection from certain regions. This law also effectively creates a "right to explanation" under which a subject may ask for an explanation of the logic or algorithm involved in reaching an automated decision. This has immediate implications for using interpretable machine intelligence algorithms in public decision making applications. In order to address the two key problems of bias in data and bias in algorithms, it is important to understand the life cycle of machine intelligence applications. We outline the life cycle of machine intelligence applications and then focus on two key points in the cycle where this regulation is most likely to introduce bias. After identification, necessary measures can be taken to address the risks introduced by the bias.Then we outline some strategies for mitigating those biases to mitigate the risks for public.

Introduction

Machine intelligence and the related fields of artificial intelligence (AI), machine learning (ML) and big data have gained widespread traction in the past decade. Machine intelligence has moved beyond its sandbox applications of number recognition, basic translators and film recommender systems. Today Machine Intelligence Systems (MISs) are widely deployed in the public sphere and are used in predictive policing (Perry 2013), legal assistance (Surden 2014; Katz 2012), insurance analytics (Rose 2016), healthcare (Araújo, Santana, and Neto 2016), advertising technology, risk and fraud detection (Levin, Pomares, and Alvarez 2016; Costello, Ianakiev, and Johnson 2015), customer intelligence, crime prevention, and terrorist activity detection (Akhgar et al. 2015). Discriminating bias is the presence of a blind spot in the logic of an MIS that disproportionately affects a certain part of the public. Bias in the decision produced by MISs can carry significant social costs and personal risks for individuals or groups who are subject to these automated decisions (Lum and Isaac 2016). With their deep integration into financial, health, and security systems the margin for bias in MIS deployment is narrowing and the need for mitigating bias risks is growing. New data protection regulations introduced by the European Union seek to address these issue by providing more power to the members of public to control their personal data and to demand explanation for the algorithms to which they are subject to, but this creates some complications of its own. The one we focus on in this paper is the introduction of bias in the collection of data sets and selection bias in algorithms used for making automated decisions. The objective of this paper is to look at it through the lens of the life cycle of machine intelligence applications, and points in the cycle where there is most risk of introduction of bias. After identification of these points, necessary measures can be taken to address the risks introduced by the bias.Then we outline some strategies for mitigating those biases to mitigate the risks for public.

First we give some background on the risks involved in machine intelligence applications, an estimation of the scale of the machine intelligence industry, followed by a brief summary of the section of the European Union regulations that we are interested in. This is followed by a brief examination of the life cycle of machine intelligence applications. Then we focus the discussion on risks introduced by bias in the data, and the risks introduced by bias in the selection of algorithms.

Background

The machine intelligence industry is currently an unregulated market with a lack of formalised risk assessment and quality standardisation. Neither academia nor the industry has the capacity for fully mitigating MIS specific risks that the public is exposed to. These risks introduced by bias in machine intelligence applications include but are not limited to automated credit or insurance coverage being wrongly denied, racial bias in profiling (Lum and Isaac 2016; Rutkin 2016), discrimination in educational or work opportunities (Chalfin et al. 2016).

In March, 2016 Microsoft launced a chatbot on Twitter, named Tay and invited people to interact with it. Malicious individuals quickly took over the process and taught Tay to repeat sexist, racist, and anti-Semitic phrases. The chat-bot was taken off-line within a day of being launched. Representatives from Microsoft admitted that even though they had "conducted extensive user studies with diverse groups ... under a variety of conditions" they had made a "critical oversight for this specific attack" (Lee 2016; Michael 2016; Ohlheiser 2016)

In 2013, a researcher from Harvard University discovered that a Google search for her name led to online advertising suggesting that arrest records for the name should be looked up (Rutkin 2016). After carrying out a follow-up study on Google AdSense it was found that searches for first names that are primarily assigned to black children were 25% more likely to generated ads suggestive of an arrest record, regardless of whether the exact full name had an arrest record in the advertising company's database or not (Sweeney 2013)

In June, 2015 a user of Google photos app found that two black people had been automatically tagged as "gorillas" using smart tagging (Alcine 2016; Tutt 2016) The photo storage website Flickr also had similar problems where the photo of a black man was tagged with "animal" (Rutkin 2016).

There is concern in the ethics and legal communities regarding the implications of bias in algorithms and the harms that they can cause (Tutt 2016). There are external calls for managing the risk and for design principles which ensure mitigation of discrimination in algorithms (Wittkower 2016). A novel research strand has emerged that seeks to raise the machine intelligence research community's awareness of the risks inherent in MISs. In papers (Sculley et al. 2014) and (Sculley et al. 2015) these risks are analysed through the lens of *technical debt*. Technical debt is the compromise that large system developers make between code that is quick to prototype in the short run vs code that is more robust in the long run. Specifically, these papers argue that MISs have a special capacity for incurring technical debt, because they have all of the maintenance problems of traditional code as well as an additional set of MIS-specific problems. MISs have some fundamental characteristics that set them apart from traditional software systems - namely their reliance on data to make decisions. While papers (Sculley et al. 2014) and (Sculley et al. 2015) certainly put forward pioneering ideas and initiatives in the area of MIS-related risks, they focus on the developer side risks rather than the public side risks. In the overall life cycle of a machine intelligence application there are some key risks to which public individuals get exposed to when MISs are deployed in public decision making. Due to the way in which a machine intelligence based decision making system is dependent on the data that it has ingested, multiple failure points arise when *bias* get introduced in the underlying data. Bias arises when the the structure and origins of data introduce errors in the final outputs of automated decision making systems. These risks are a matter of great concern public members who are not involved in the underlying research and development process.

Risks associated with the utilisation of products in the pharmaceutical, financial and environmental sectors have been extensively communicated to all their stakeholders. However, the same cannot be said of the machine intelligence technology industry. The absence of guidelines to assess and communicate product and service risks in MISs can lead to compromised trust between developers of these systems and the public that are affected by the actions of the automated decision making processes. There are a growing number of companies in the MIS ecosystem which are offering a variety of services to the public independently. The exposure of individuals to risk is compounded because each individual uses and is subject to multiple MISs. The financial risks associated with the use of MISs are currently not addressed in literature and majority of public don't have access to in house expertise on MISs. Many industries and government departments can suffer from an information asymmetry to comprehensively assess the risks of incorporating MISs in their decision making pipeline because no unified quality metrics or reliability testing guidelines exist to validate the accuracy of MIS implementation. Traditional methods used in software systems' testing are currently also used as the de facto industry standard in MLS implementation testing. This however, fully ignores the differences between MISs and traditional software systems, i.e. the ability of MISs to learn from data and their reliance on probabilistic methods.

New protection laws that have been introduced by the European Union under the General Data Protection Regulations try to address these risks by offering the individuals subject to machine intelligence and automated decision making systems to have more control over their personal data and to offer them the right to demand an explanation for the underlying logic used in a automated decision making process. We summarise these in a following section of the paper. This has direct implications for machine intelligence applications that rely on large scale collection of data and advanced algorithms with limited interpretability. As more people opt out of data collection services under these regulations, there is the distinct possibility that unexpected biases will develop in the collected data sets. If not addressed, these biases can further propagate through the decision making process and further exacerbate discrimination for certain members of public. In the same vein, attempting to avoid complex and nuanced algorithms in order to maintain simplicity and interpretablity, in order to comply with these regulations, could result in a shift towards algorithms that are simple to explain but which don't fully provide the benefits that are possible from machine intelligence. Simpler algorithms could prove too blunt to fully capture and address the complexity of decision mak-

ing required in the public space and may end up heavily discriminating against certain factions of public. To gauge the necessity of mitigating these risks we provide a brief assessment of the scale of the machine intelligence industry.

The extent of the machine intelligence industry as an indicator of the associated risk

There has been widespread excitement about machine intelligence applications over the last decade. The excitement is shared across the media, venture capitalists, corporates and the general public at large. The confluence of advancements in large scale data management, computing power and algorithmic sophistication has dramatically accelerated high-performance outcomes with the result that many providers of machine intelligence applications and automated decision making products have become part of the MIS eco-system.

The aforementioned trend is reflected by the financing history of companies that are offering machine intelligence applications. Data on the companies in this sector is showing large scale growth in activity. Over 60 companies raised equity funding rounds in 2015, up from 8 in 2011, amounting to a nearly sevenfold increase in deal activity. The amount invested in start-ups rose to $318 million in 2015; up more than 10-fold from $25 million in 2011 (CB Insights 2016). In 2015, new machine intelligence focused business received 5% of all venture capital funding, with investments going to companies in 13 separate countries and 10 industry categories, including business intelligence, e-commerce, and healthcare. As of the fourth quarter of 2016, there are around 1450 machine intelligence companies, across 73 countries with a total of $8.5 billion in funding (Venture Scanner Insights 2016). These businesses are primarily focused on developing technologies around the core areas of image processing, natural language processing, deep learning and predictive analytics. When we combine these investment figures with the number of public members directly using the services of these businesses we can see that their is a proliferation of risk associated with machine intelligence applications.

As the deployment and complexity of MISs and automated decision services increases the problems of bias risks, discussed in the previous section, will only escalate. In order to address the two key problems of bias in data and bias in algorithms, it is important to understand the life cycle of machine intelligence applications. This will help to identify where bias enters the life cycle so that necessary measures can be taken to address the risks introduced by the bias. There is a gap between the extent of risks that threaten MIS functionality and the measures currently available to address those risks. This seemingly obvious gap provides an opportunity for machine intelligence researchers to take a lead in the design of data and algorithm frameworks which avoid discrimination and bias, yet offer nuance for advanced decision making and interpretabiltiy.

Right to object and automated individual decision making

In April 2016, the European Parliament adopted a set of data protection laws for the collection, storage and processing of personal data, the General Data Protection Regulations (GDPR) (The European Parliament and The European Council 2016). While most of the regulations outlined are concerned with how data collection and storage, we focus on *Section 4: Right to object and automated individual decision making*. An excerpt from this is shown in the inset. This section introduces important concepts that have far reaching consequences for data analysis and automated decision making and as such could prohibit a wide range of algorithms currently in use in machine intelligence applications.

Section 2 - Article 15
Right of access by the data subject

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

   ...

   (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

Section 4 - Article 21
Right to object

1. The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions. The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims.

   ...

6. Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.

Section 4 - Article 22
Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

   ⋮

A more through discussion of this section of GDPR can be found in (Goodman and Flaxman 2016). We summarize some key terms that have been used in the GDPR in *Article 4: Definitions*:

data subject: "an identifiable natural person is one who can be identified, directly or indirectly, in particular by refer-

ence to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

personal data: "means any information relating to an identified or identifiable natural person".

processing: "means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

profiling: "means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".

As might be inferred, the language of these regulations poses significant challenges to current practices in machine intelligence applications. The *right to object*, when practiced widely will have the effect of introducing unexpected biases in data as sections of public will be missing from data samples. This in turn increases the risks of discrimination towards public members who are not well represented in the data. The *right to meaningful information about the logic involved* is in essence a "right to explanation" (Goodman and Flaxman 2016), and has implications for the interpretablity of machine intelligence algorithms that are deployed in order to arrive at the automated decision. Such interpretability may have to come at the cost of performance of the machine intelligence application. As the field stands, more interpretable and easier to convey models might lack the nuance needed to handle complex data for large scale applications, therefore biases might be introduced in the automated decision making process. This in turn exacerbates the risk of discrimination against certain individuals of the public, especially ones on whom data might be scant in the first place, due to them being minorities in the larger population.

We now summarise the life cycle of machine intelligence applications in order to place where these two bias risks arise in the life cycle.

Life cycle of machine intelligence systems

MISs as applied in public space setting are not just a standalone entity, rather they constitute a life cycle of the machine intelligence algorithm and the data that feeds this algorithm. Machine intelligence applications present a unique challenge in that they are not just software that statically fits into an existing infrastructure - rather they are a dynamic component that, when implemented properly, interact with all aspects of the infrastructure which it occupies and alter the very processes that they form part of. In this paper the infrastructure that we are interested in are ones where members of the public are affected by the automated decisions of machine intelligence applications. A good understanding of the MIS life cycle is necessary for all individuals involved in the development, deployment and public use of machine intelligence applications (Flyvbjerg, Glenting, and Rönnest 2004).

The MIS life cycle has two closely interacting parts, the cycle of the machine intelligence algorithm and the cycle of the data that is used with this algorithm. Machine intelligence algorithms, whether custom designed or off the shelf solutions, interact deeply with initially collected data, they harness insights from the data and they steer the collection of new data. New data in turn drives the selection of new machine intelligence algorithms and the cycle goes through its next iteration.

In order to capture this interaction of machine intelligence algorithms and data, a nested product life cycle is shown in Figure 1. Both the algorithm and data cycles closely mirror each other at each stage of the process and are enmeshed in their practical implementation. An explanation of each cycle is given below.

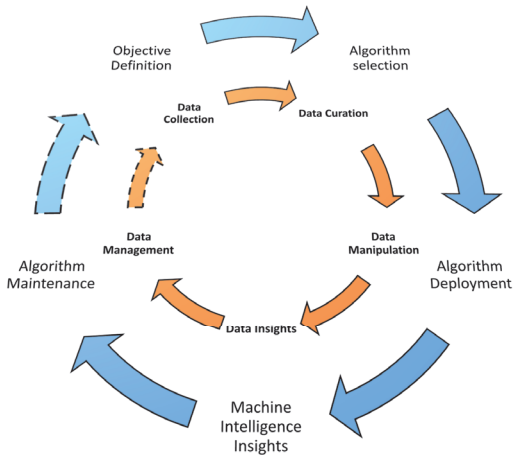

Figure 1: Machine intelligence system life cycle: algorithm cycle outside in blue and data life cycle inside in orange

Machine intelligence algorithm life cycle

The Machine intelligence algorithm cycle starts off with an objective definition phase brought about by a need for automating a decision making process. This step in itself remains subject to revision and refinement as the experts schooled in the language and methods of machine intelligence interact with members of public or other sources of feedback which give clarity on the problem being solved.This phase is important in terms of expectation management and setting up realistic goals for the various improvements possible from deploying MISs.

After an objective has been identified, a relevant machine intelligence algorithm is selected from the many possible candidates currently available from research. This is a first key point where discrimination bias may enter the life cycle, as different algorithms have different strengths when faced with different types of data, therefore as the nature of the underlying data changes, an algorithm which works for one

public body may not prove appropriate for another public body.

Once an appropriate algorithm has been selected, it is then deployed into the larger infrastructure. This infrastructure can be any system or functioning entity which requires the services of automated decision making being provided by the machine intelligence algorithm. The deployment phase is crucial in terms of ensuring long term robustness of the MIS, since improper integration of the algorithm within the larger infrastructure may result in flawed decision making.

After the algorithm has been incorporated into the infrastructure real time insights can be gained. These insights may come in the many forms such as intelligent analysis of historical data or in the form of predictions for future trajectories, or in the form of advice regarding business targets etc. It is critical at this phase of the algorithm life cycle that the insights are interpreted properly. Interpretation of insights requires people or teams with cross discipline skills in data science, machine intelligence, experiment design, and decision making in order to re-frame the complex outputs of the machine intelligence algorithm into actionable insights from which public can benefit.

The successful harnessing of machine intelligence insights must be followed by consistent maintenance of the algorithm especially by reducing technical debt, and system updates that keep the algorithms secure from malicious tampering, as well as updates to the data that may be helpful in keeping the insights relevant. Maintenance of the machine intelligence algorithm should also encapsulate a continuous assessment of whether the machine intelligence insights are still in line with expectations, so sanity checks must be deployed to consistently monitor the algorithm life cycle.

The insights obtained from the algorithm cycle will give rise to new questions and challenges which will inform the next iteration of objective setting.

Machine intelligence data life cycle

Closely entwined with the machine intelligence algorithm life cycle is the life cycle of the data. Access to large scale reliable data is required to effectively harness insights from machine intelligence algorithms. Data is dynamic in that it keeps changing throughout the various phases of the cycle and so it must always be kept in mind how those changes may affect the outcomes when the algorithm cycle interacts with the data.

The first step in the data life cycle is the collection of data. During the collection of data the sources and types of data being collected should be closely monitored and checked for bias against different members of the public. Data origins are a key source of discrimination bias in the MIS life cycle. If certain populations of public are not represented adequately in the data sources then it is highly likely that any automated decision making process that is implemented will have bias built into it and will put those populations at undue risk of discrimination. We will discuss this in further detail in the following section.

After the collection of data, the focus shifts to data curation i.e. identifying gaps in the collected data, and determining the resources required to fill those gaps. Depending on the nature of the gaps in the data, bias correction strategies should be deployed in the algorithm selection so that gaps in the data can be effectively addresses and the effects of these gaps are mitigated. If these gaps are not well-addressed then discriminating biases can propagate through the cycle, exposing members of the public to risk.

Data manipulation entails making the data fit for interaction with the machine intelligence algorithm by putting the data in the correct format, cleaning it of spurious values and making sure that the data set has the requisite characteristics as required by the algorithm. Data manipulation is the step in which the data comes into direct contact with the deployed algorithm in order to produce intelligent insights in line with the pre determined objectives.

Data insights are closely related to the concept of the algorithm insights but differ in one key aspect. Data insights may also reveal the deficiencies of the data, in addition to the relevant intelligent analysis and predictions. Insights about blind spots in the data may require inbuilt checks from the MIS and must be actively pursued in order to ensure robustness of the overall life cycle.

Data management and maintenance is a key step to sustain the health of the algorithm and data life cycle, without this step being actively carried out the quality and trustworthiness of any MIS will deteriorate and cause expectation mismatches over the long term. Data management will involve ensuring that the data is up to date, the characteristics of the data are still in line with the assumptions of the machine intelligence application and the discontinued data streams are not part of the data corpus.

Based on good data management, new data will be created which will feed the next iteration of the data life cycle.

Discrimination risk in the origins of data

Bias-In-Bias-Out risk

Bias-In-Bias-Out (BIBO), also known as Garbage-In-Garbage-Out (GIGO), is one of the primary risks for public members subject to machine intelligence applications. The acronym BIBO is used to express the idea that incorrect and poor-quality input with inherent biases will produce faulty and discriminating outputs in data analysis processes (Bininda-Emonds et al. 2004). Any MIS, no matter how sophisticated, has the potential of being affected by BIBO because MISs use data-driven inputs to calibrate their decision making processes, so biased data will result in biased calibrations of the algorithm (Barga, Fontama, and Tok 2015).

In many cases, the performance of the machine intelligence algorithm if properly selected and implemented, will outperform that of a human expert. However, if the provided data are erroneous, contaminated or biased for discrimination, the machine intelligence algorithm is unlikely to be able to overcome it. MISs just cannot perform ‘data alchemy’ and turn data lead into gold (Barga, Fontama, and Tok 2015). This risk is universal to all data processing systems.

Due to the GDPR data protection laws, data sets that are collected at large scale will contain some inherent biases due to the missing data points on some fractions of the public. This means that utilization of this biased data without

proper curation will result in unintentional discrimination risks against certain members of public. Therefore it is imperative to manage this data bias.

Managing the bias risk in data collection

The key point at which BIBO risk enters the MIS life cycle is at the point of data collection. In order to manage the risk introduced by the origins of the data, the provenance of data must be analysed critically i.e. is the data sourced from a third party or was is collected in-house (Simmhan, Plale, and Gannon 2005; Buneman, Khanna, and Wang-Chiew 2001). If it was sourced in-house then the assumptions underpinning the type of data collected, the geographical region it was collected in and the frequency it was collected at needs to be understood and aligned with the objectives of the machine intelligence application. It might be that the data was sourced from Asia but completely unrepresentative of the Americas and therefore will produce discriminatory outputs when fed into a machine intelligence algorithm. Sourcing the wrong type of data from the wrong region at the wrong time of year will result in GIGO risk. Knowing the provenance of the data will enable the management of bias risks introduced during the collection of data and allow much better curation of data. For managing the quality of data, standards such as the ISO 8000 *Data Quality* standard (Benson 2008; Gitzel, Turring, and Maczey 2015) should be put into practice across the whole life cycle of the MLS.

Selection risk in interpretability of algorithms

The black-box risk

Research in the field of machine intelligence currently out paces our ability to amply explain the underlying technical workings of the algorithms being used. This technical complexity combined with the scale at which machine intelligence algorithms need to be deployed in order to harness their fully use introduces a deep opacity in the MIS life cycle (Burrell 2016). Often times machine intelligence applications are part of large scale public works which need some measure of secrecy in order to protect them from malicious actors. This is compounded by the fact that the threshold of technical literacy required for understanding most machine intelligence techniques creates a barrier to entry and produces another layer of opacity. These factors combine to produce the black-box effect, whereby machine intelligence algorithms are used as part of public decision making processes without being fully understood by the majority of the members of public being directly affected by these decisions. When machine intelligence applications have social consequences the opacity in machine intelligence algorithms can be quite damaging due to its discriminatory effects (Lum and Isaac 2016; Tutt 2016)

Since the new regulations laid out in GDPR effectively introduce the "right to explanation" (Goodman and Flaxman 2016), this posses a challenge as many of the practically deployed machine intelligence algorithms are not easily interpretable. There is a risk that in order to avoid the challenge machine intelligence applications will shift towards using more easily interpretable algorithms. The challenge with simpler and more interpretable algorithms would be that they may lack the nuance of scale and technical rigour that drove the development of more advanced techniques in the first place. Therefore these more interpretable algorithms may carry the risk of introducing a discrimination bias in the system by having a larger margin of error, and this error will have a larger significance for disadvantaged minorities within the public body.

Managing the bias risk in algorithm selection

The lack of interpretability fundamentally arises from an informational asymmetry between those developing and deploying the machine intelligence algorithms and the ones who are directly subject to the decisions of the algorithms. Information asymmetry between MIS developers and MIS subjects creates a hight risk situation (Akerlof 1995). The algorithm developers will have a deeper understanding of the technical aspects of the objective, whereas public subjects may have a deeper understanding of the daily realities and consequences of the automated decisions. The task of effectively communicating how and what types of data manipulation is applied before the results are produced can be quite difficult due to the technical opacity involved (Budzier and Flyvbjerg 2012).

In order to tackle the risks arising from the opacity of machine intelligence applications the public subjects interacting with the algorithms should be provided with opportunities to gain at least a broad level understanding of the MIS being implemented. Rather than being treated as a technology project, machine intelligence projects should be treated as a social paradigm, with MIS deployment teams spending time to gain a deep understanding of public members who will be directly affected by the automated decisions. Many online demos now exist to visually explain the different processes that machine intelligence algorithms are performing. A complete MIS may utilise many of these algorithms there by compounding complexity, but explaining these different components should be part of the service level agreement (SLA) established between the MIS developers and public subjects(Flyvbjerg 2008). By overcoming informational asymmetry discrimination biases in algorithm selection can be mitigated.

Conclusion

The use of machine intelligence systems (MISs) carries the risk of discriminating against certain members of public when bias is introduced into the MIS life cycle at the point of data collection and algorithm selection. There can be deep social implications of these discriminating biases. New General Data Protection Regulations introduced by European Union, have direct consequences for introducing data bias if decision makers circumvent this law by avoiding data collection from certain regions. This law also effectively creates a "right to explanation" under which a subject may ask for an explanation of the logic or algorithm involved in reaching an automated decision. This has immediate implications for using interpretable machine intelligence algorithms in public decision making applications. Closely managing the origins

of data and bias reduction via curation can decrease risks of discrimination in data collection, while reducing information asymmetry with the public can help reduce the risks introduced during algorithm selection.

References

Akerlof, G. 1995. *The market for "lemons": Quality uncertainty and the market mechanism*. Springer.

Akhgar, B.; Saathoff, G.; Arabnia, H.; Hill, R.; Staniforth, A.; and Bayerl, P. 2015. *Application of Big Data for National Security: A Practitioner's Guide to Emerging Technologies*. Butterworth-Heinemann. Elsevier Science & Technology Books.

Alcine, J. 2016. "Google Photos, y'all fucked up. My friend's not a gorilla.". *Tweet*. Available at: https://twitter.com/jackyalcine/status/615329515909156865.

Araújo, F. H. D.; Santana, A. M.; and Neto, P. d. A. S. 2016. Using machine learning to support healthcare professionals in making preauthorisation decisions. *International Journal of Medical Informatics* 94:1–7.

Barga, R.; Fontama, V.; and Tok, W. H. 2015. Data Preparation. In *Predictive Analytics with Microsoft Azure Machine Learning*. Springer. 45–79.

Benson, P. 2008. ISO 8000 the International Standard for Data Quality. *MIT Information Quality Symposium* July 16-17(2008).

Bininda-Emonds, O. R. P.; Jones, K. E.; Price, S. A.; Cardillo, M.; Grenyer, R.; and Purvis, A. 2004. Garbage in, garbage out. In *Phylogenetic supertrees*. Springer. 267–280.

Budzier, A., and Flyvbjerg, B. 2012. Overspend? Late? Failure? What the data say about IT project risk in the public sector. *Commonwealth Governance Handbook* 13:145–157.

Buneman, P.; Khanna, S.; and Wang-Chiew, T. 2001. Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001*. Springer. 316–330.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1).

CB Insights. 2016. Deep Interest In AI: New High In Deals To Artificial Intelligence Startups In Q4'15. https://www.cbinsights.com.

Chalfin, A.; Danieli, O.; Hillis, A.; Jelveh, Z.; Luca, M.; Ludwig, J.; and Mullainathan, S. 2016. Productivity and Selection of Human Capital with Machine Learning. *The American Economic Review* 106(5):124–127.

Costello, T.; Ianakiev, K. G.; and Johnson, J. 2015. Systems and Methods for Computerized Fraud Detection Using Machine Learning and Network Analysis. US Patent 20,160,117,778. US Patent App. 14/921,773.

Flyvbjerg, B.; Glenting, C.; and Rönnest, A. K. 2004. Procedures for dealing with optimism bias in transport planning. *London: The British Department for Transport, Guidance Document*.

Flyvbjerg, B. 2008. Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European planning studies* 16(1):3–21.

Gitzel, R.; Turring, S.; and Maczey, S. 2015. A Data Quality Dashboard for Reliability Data. In *2015 IEEE 17th Conference on Business Informatics*, volume 1, 90–97. IEEE.

Goodman, B., and Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY*.

Katz, D. M. 2012. Quantitative Legal Prediction-or-How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry. *Emory Law Journal* 62.

Lee, P. 2016. Learning from Tay's introduction. *Official Microsoft Blog* March 25.

Levin, I.; Pomares, J.; and Alvarez, R. M. 2016. Using Machine Learning Algorithms to Detect Election Fraud. *Computational Social Science* 266.

Lum, K., and Isaac, W. 2016. To predict and serve? *Significance* 13(5):14–19.

Michael, K. 2016. Science Fiction Is Full of Bots That Hurt People: ...But these bots are here now. *IEEE Consumer Electronics Magazine* 5(4):112–117.

Ohlheiser, A. 2016. Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac. *The Washington Post* 25.

Perry, W. L. 2013. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Rose, S. 2016. A Machine Learning Framework for Plan Payment Risk Adjustment. Wiley Online Library.

Rutkin, A. 2016. Is tech racist? The fight back against digital discrimination. *New Scientist* 231(3084):18–19.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; and Young, M. 2014. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, 2503–2511.

Simmhan, Y. L.; Plale, B.; and Gannon, D. 2005. A survey of data provenance in e-science. *ACM Sigmod Record* 34(3):31–36.

Surden, H. 2014. Machine Learning and Law. *Washington Law Review* 89(1):87–1467.

Sweeney, L. 2013. Discrimination in online ad delivery. *Communications of the ACM* 56(5):44–54.

The European Parliament, and The European Council. 2016. General Data Protection Regulation. *Official Journal of the European Union* 2014(April):20–30.

Tutt, A. 2016. An FDA for Algorithms. *Administrative Law Review* 67.

Venture Scanner Insights. 2016. Artificial Intelligence Market Overview – Q4 2016.

Wittkower, D. E. 2016. Principles of anti-discriminatory design. *Ethics in Engineering, Science and Technology (ETHICS), 2016 IEEE International Symposium on* 1–7.