Preface

More than ever before, data, information, algorithms and systems have the potential to influence and shape our experiences and views. With increased access to digital media and the ubiquity of data and data-driven processes in all areas of life, an awareness and understanding of topics, such as algorithmic accountability, transparency, governance and bias, are becoming increasingly important. Recent cases in the news and media have highlighted the wider societal effects of data and algorithms requiring we pay it more attention.

The BIAS workshop brought together researchers from different disciplines who are interested in analysing, understanding and tackling bias within their discipline, arising from the data, algorithms and methods they use. The workshop attracted 14 submissions, including research papers and extended abstracts. After a peer reviewing process in which each submission received three independent reviews, the following six papers were accepted and are included in these proceedings:

- Claude Draude, Goda Klumbyte and Pat Treusch: *Re-Considering Bias: What Could Bringing Gender Studies and Computing Together Teach Us About Bias in Information Systems?*
- Christoph Hube, Besnik Fetahu and Robert Jäschke: *Towards Bias Detection in Online Text Corpora*
- Vasileios Iosifidis and Eirini Ntoutsi: *Dealing with Bias via Data Augmentation in Supervised Learning Scenarios*
- Serena Oosterloo and Gerwin van Schie: *The Politics and Biases of the "Crime Anticipation System" of the Dutch Police*
- Alan Rubel, Clinton Castro and Adam Pham: *Algorithms, Bias, and the Importance of Agency*
- William Seymour: *Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?*

The papers cover a wide range of research topics: from conceptual discussions of algorithmic transparency and fairness to empirical research and case studies.

William Seymour (page 2) discusses the relationship between the fairness of an algorithm and its transparency and the important distinction between process transparency and output transparency. For most effective machine learning algorithms we cannot hope to obtain process transparency, as their inner workings are beyond conscious human reasoning. Seymour argues that a viable alternative is to analyse the transparency of the outcome of an algorithm. He also presents two exemplary methods – local explanations and statistical analysis – that could help to understand the fairness of the outputs.

Alan Rubel, Clinton Castro and Adam Pham (page 9) address the notions of agency and autonomy with regard to algorithmic systems. While debates about biases in algorithmic systems often emphasise potential and actual harms, the

authors argue that our concerns about algorithms should not be limited to such issues. Moving the debate forward beyond interest in algorithmic harms, they argue that the "moral salience" of algorithmic systems cannot be understood without also addressing their impacts on human agency, autonomy and respect for personhood.

Claude Draude, Goda Klumbyte and Pat Treusch (page 14) explore the potential for theoretical frameworks from gender studies – including Haraway's "situated knowledges" and Harding's "standpoint theory" – to inform a better understanding of how bias emerges within information systems. With a particular focus on issues of androcentrism, over/underestimation of gender differences and the stereotyping of gender traits in the workings of information systems, their paper considers how feminist insights might help to account for and prevent bias in information system design.

Christopher Hube, Robert Jäschke and Besnik Fetahu (page 19) present a method for identifying language bias within textual corpora using word embeddings, based on word2vec. This includes a two-stage process in which firstly seed words indicating bias are extracted from Conservapedia, a dataset that includes opinionated political articles. The second step uses word2vec to identify bias words involving the seed list created previously. The approach iterates to keep growing the list of bias words that could be used to form feature vectors for tasks such as supervised learning.

Vasileios Iosfidis and Eirini Ntoutsi (page 24) describe techniques for data augmentation (SMOTE and oversampling) to deal with cases of class imbalance where under-represented groups can affect data-driven methods, such as supervised learning. Their experiments on the Census Income and German Credit datasets show that the classes can be more equally represented using data augmentation without affecting overall classification performance. This is particularly important when dealing with biases in datasets around certain attributes, such as gender and race, where the methods proposed in the paper can reduce classification errors for potentially discriminated groups.

Serena Oosterloo and Gerwin van Schie (page 30) walk us through a crime prediction system currently being used in the Netherlands, from a critical data studies perspective. Their paper illustrates various sources of inaccuracies in the system, including those that cannot be helped – because the necessary attribute cannot be measured with great precision in the offline world – as well as those that result from human biases (e.g., the choices made during the process of classifying a crime and the parties involved).

The workshop was opened by a keynote from *Ansgar Koene*, University of Nottingham, (page 1) discussing socio-technical causes of bias in algorithms and systems and the role of policies and ethical standards.

Sheffield, March 25, 2018
`https://ir.shef.ac.uk/bias/`

Jo Bates
Paul D. Clough
Robert Jäschke
Jahna Otterbacher