Using SHAP Values for Clustering-based Bias Detection

Mirthe Dankloff[1,2], **Emma Beauxis-Aussalet**[1, 2], **Jacco van Ossenbruggen**[1, 2]
[1]Civic AI Lab
[2]Vrije Universiteit Amsterdam
NU building, De Boelelaan 1111
1081 HV Amsterdam, The Netherlands
m.e.dankloff@vu.nl, jacco.van.ossenbruggen@vu.nl, e.m.a.l.beauxisaussalet@vu.nl

Abstract

A key form of AI bias occurs when a model produces higher error rates for specific demographic groups based on (sensitive) features such as gender or ethnicity. Explanation methods like SHAP are used to detect biases in AI models. However, it can be unknown beforehand for which features a model exhibits higher error rates or which SHAP values are most informative for detecting such bias. The Hierarchical Bias Aware Clustering (HBAC) algorithm is an unsupervised method designed to identify clusters with higher error rates using any combination of features. We study whether including SHAP values as input features to HBAC can be beneficial for bias detection. We explore the HBAC results by combining different sets of features including SHAP values, error labels, regular and sensitive features. For each combination, we evaluate the HBAC results with three quality criteria: the clusters should have: (i) distinctive error rates, (ii) distinctive sensitive features, and (iii) general cluster separability. Using the COMPAS dataset, we compare three clustering methods that can be embedded within HBAC: K-means, K-Prototypes, and DBSCAN. Our study shows that when SHAP values are excluded from the HBAC input feature combinations, the identified clusters have high quality across all criteria and clustering methods. When SHAP values are included, the general cluster separability largely decreases for most of the feature combinations, while the other two criteria seldom improve. Furthermore, clustering solely on SHAP values yields poor results for all quality criteria. These results raise questions about the usefulness of SHAP values as indicators for AI bias detection. Future work must further investigate such potential limitation of SHAP values, which might not be suitable for post-hoc bias analysis.

Introduction

Bias and discrimination occurs when machine learning models systematically contain more errors for individuals or groups of people based on their sensitive features such as ethnicity, age, or gender (Buolamwini and Gebru 2018). Removing sensitive features from a model is often insufficient to mitigate bias due to correlations between sensitive features and other (proxy) features (Helwegen, Louizos, and Forré 2020; Barocas, Hardt, and Narayanan 2019). Moreover, intersectional harms occur, such as specific forms of discrimination for members of multiple sensitive features. Defining which features to include in a model can thus be challenging as discriminated groups can be unexpected, unlabelled and undefined depending on the use case at hand (Luong, Ruggieri, and Turini 2011; Misztal-Radecka and Indurkhya 2021; Chakraborty, Peng, and Menzies 2020).

Interpretability of a machine learning model is important for evaluating bias and discrimination (Doshi-Velez and Kim 2017). To this end, post-hoc explanation methods like SHAP are employed to interpret which features indicate bias in a model (Silberg and Manyika 2019; Jain, Ravula, and Ghosh 2020). SHAP is a model-agnostic explanation method designed to interpret model predictions by assigning each input feature an importance value given a set of features (Molnar 2023; Lundberg and Lee 2017). It allows for a local or global interpretation, looking at a specific data point or at the average feature importance for the whole test set (Molnar 2023). Despite its widespread use, SHAP faced criticism for not offering human-friendly and trustworthy explanations to end users (Miller 2019; Kumar et al. 2020; Bhatt et al. 2020). Moreover, prior studies have shown that SHAP is unreliable when it comes to interpreting bias in a model (Slack et al. 2020; Chakraborty, Peng, and Menzies 2020; Dimanov et al. 2020).

In contrast, unsupervised clustering methods have proven effective for detecting discriminatory bias when the specific combination of features leading to bias are unknown (Luong, Ruggieri, and Turini 2011; Misztal-Radecka and Indurkhya 2021; Nasiriani et al. 2019; Muhammad 2021). Clustering methods partition a dataset into clusters such that instances in the same cluster are more similar to each other than to those in other clusters (Ji et al. 2012; Huang 1997). Hidden patterns of bias can be discovered by clustering on negative and positive decision outcomes for similar instances (Nasiriani et al. 2019), deviation in model performance (Misztal-Radecka and Indurkhya 2021), and based on higher and lower error rates (Muhammad 2021). The latter refers to the Hierarchical Bias Aware Clustering (HBAC) Algorithm, which forms the basis for this work.

In this paper, we investigate whether can SHAP values be used as an indicator of the potential bias in machine learning results. To address this problem, we investigate how SHAP values are related to the groups of data points that exhibit high error rates. We use HBAC to identify such groups of data points with high error rates. Applied to a classification problem, we investigate whether using SHAP values as input features can help identifying the clusters of classification errors. We research how using SHAP values improves the HBAC clustering-based bias detection through three **cluster quality criteria**:

- Separability of errors: The clusters' differences in probability of error should be the largest s.
- Separability of social groups: The clusters' differences in sensitive features should be the largest.
- General separability: The clusters' inter- and intra-cluster variance (e.g. silhouette scores) should be the most optimal.

Thus, we investigate the following research questions:

- **RQ1**: What are the cluster quality differences when SHAP values are used as input features of the HBAC bias detection?
  - 1-1: How does using SHAP values as input features of the HBAC bias detection impact the detection of clusters with higher error rates?
  - 1-2: How does using SHAP values as input features of the HBAC bias detection impact the detection of clusters that are representative of social groups with specific sensitive features?
  - 1-3: How does using SHAP values as input features of the HBAC bias detection impact the general cluster separability?

Besides adding SHAP values as input features, the HBAC algorithm might also be optimised by using (i) alternative clustering techniques embedded within HBAC (Muhammad 2021), (ii) error labels as input features (Muhammad 2021) (e.g., 0 for correct classification, 1 for errors), or (iii) sensitive features only (e.g., to focus on vulnerable social groups). Therefore, we also address the following research questions:

- **RQ2**: How does the clustering quality differ when using K-means, K-Prototypes, or DBSCAN within HBAC?
- **RQ3**: How does the clustering quality differ when adding error labels as an input feature of HBAC?
- **RQ4**: How does the clustering quality differ when using only sensitive features, with or without their SHAP values, as input features of HBAC?

We answer these research questions by experimenting with different experimental conditions that contain different sets of input features (Table 1)

To answer RQ1 and the SQ's, we compare clustering outcomes for data frames with (i) regular features, and SHAP values, (ii) only regular features, and (iii) only SHAP values. The cluster quality is compared for K-means, K-Prototypes and DBSCAN clustering methods and are tested for significance. To answer RQ2 and RQ3, we take Error values as input features and compare the clustering outcomes for data frames with (i) regular features, SHAP values, and Error labels, (ii) only SHAP values and Error labels and (iii) Regular features and Error labels.

Related Work

Prior work has been done on clustering methods to detect discrimination through bias. Nasiriani et al.(Nasiriani et al. 2019) used top-down hierarchical clustering to detecting positive and negative bias for statistically similar data points. Individuals with similar characteristics are compared for either receiving a positive or negative decision values. For each identified cluster, the level of discrimination is the difference in ratio between positive and negative decisions. Finally, they investigate whether certain clusters receive more positive than negative decision labels. Misztal-Radecka & Indurkhya (Misztal-Radecka and Indurkhya 2021) proposed the Bias-Aware Hierarchical K-means Clustering (BAH-KM) algorithm to detect negative bias in recommender systems. Negative bias in this case is understood as the deviation in model performance over clusters. The algorithm splits on those clusters that have a higher deviation compared to other clusters as it indicates that certain groups are favoured or disfavoured. The BAH-KM is model agnostic, meaning that it is suitable for a wide range of algorithms. Muhammad introduced the Hierarchical Bias-Aware Clustering (HBAC) which is an extension of the BAH-KM algorithm. They modified the BAH-KM algorithm by applying it to a classification algorithm and therefore dealt differently in how they calculated the negative bias. To this effect, they used accuracy as an evaluation metric. Moreover, they included the errors of the classification algorithm as one of the attributes and compared K-means, DBSCAN and Meanshift for finding discriminated clusters. Explanations have been framed as an important mechanism for better and fairer human-AI decision-making. [paraphrase] but research lacks the evidence on how explanations actually effect on people's perception of fairness. [cite Schoeffer et al] Previous work has shown that post hoc explanation methods like SHAP can be used to make misleading interpretations which can conceal biases in ML models (Slack et al. 2020).

Our work continues with cluster based bias detection. Contrary to previous work, we incorporate SHAP values as clustering attributes to find out how this adds to the detecting bias through cluster quality differences. Also, we look at classification systems instead of recommender systems and compare two methods that can be incorporated in the clustering framework: K-means and DBSCAN. Moreover, we extend the HBAC algorithm by incorporating true postives, false negatives, true negatives and false positives as clustering attributes instead of only one attribute. We define bias as the difference in error rate for a cluster compared to the mean absolute error rate over all clusters.

Formatting Requirements in Brief

We need source and PDF files that can be used in a variety of ways and can be output on a variety of devices. The design and appearance of the paper is strictly governed by the aaai style file (aaai24.sty). **You must not make any changes to the aaai style file, nor use any commands, packages, style files, or macros within your own paper that alter that design, including, but not limited to spacing, floats, margins, fonts, font size, and appearance.** AAAI imposes requirements on your source and PDF files that must be followed. Most of these requirements are based on our efforts to standardize conference manuscript properties and layout. All papers submitted to AAAI for publication will be recompiled for standardization purposes. Consequently, every paper submission must comply with the following requirements:

- Your .tex file must compile in PDFLaTeX — (you may not include .ps or .eps figure files.)
- All fonts must be embedded in the PDF file — including your figures.
- Modifications to the style file, whether directly or via commands in your document may not ever be made, most especially when made in an effort to avoid extra page charges or make your paper fit in a specific number of pages.
- No type 3 fonts may be used (even in illustrations).
- You may not alter the spacing above and below captions, figures, headings, and subheadings.
- You may not alter the font sizes of text elements, footnotes, heading elements, captions, or title information (for references and mathematics, please see the limited exceptions provided herein).
- You may not alter the line spacing of text.
- Your title must follow Title Case capitalization rules (not sentence case).
- LaTeX documents must use the Times or Nimbus font package (you may not use Computer Modern for the text of your paper).
- No LaTeX 209 documents may be used or submitted.
- Your source must not require use of fonts for non-Roman alphabets within the text itself. If your paper includes symbols in other languages (such as, but not limited to, Arabic, Chinese, Hebrew, Japanese, Thai, Russian and other Cyrillic languages), you must restrict their use to bit-mapped figures. Fonts that require non-English language support (CID and Identity-H) must be converted to outlines or 300 dpi bitmap or removed from the document (even if they are in a graphics file embedded in the document).
- Two-column format in AAAI style is required for all papers.
- The paper size for final submission must be US letter without exception.
- The source file must exactly match the PDF.
- The document margins may not be exceeded (no overfull boxes).
- The number of pages and the file size must be as specified for your event.
- No document may be password protected.
- Neither the PDFs nor the source may contain any embedded links or bookmarks (no hyperref or navigator packages).
- Your source and PDF must not have any page numbers, footers, or headers (no pagestyle commands).
- Your PDF must be compatible with Acrobat 5 or higher.
- Your LaTeX source file (excluding references) must consist of a **single** file (use of the "input" command is not allowed.
- Your graphics must be sized appropriately outside of LaTeX (do not use the "clip" or "trim" command) .

If you do not follow these requirements, your paper will be returned to you to correct the deficiencies.

What Files to Submit

You must submit the following items to ensure that your paper is published:

- A fully-compliant PDF file.
- Your LaTeX source file submitted as a **single** .tex file (do not use the "input" command to include sections of your paper — every section must be in the single source file). (The only allowable exception is .bib file, which should be included separately).
- The bibliography (.bib) file(s).
- Your source must compile on our system, which includes only standard LaTeX 2020 TeXLive support files.
- Only the graphics files used in compiling paper.
- The LaTeX-generated files (e.g. .aux, .bbl file, PDF, etc.).

Your LaTeX source will be reviewed and recompiled on our system (if it does not compile, your paper will be returned to you. **Do not submit your source in multiple text files.** Your single LaTeX source file must include all your text, your bibliography (formatted using aaai24.bst), and any custom macros.

Your files should work without any supporting files (other than the program itself) on any computer with a standard LaTeX distribution.

Do not send files that are not actually used in the paper. Avoid including any files not needed for compiling your paper, including, for example, this instructions file, unused graphics files, style files, additional material sent for the purpose of the paper review, intermediate build files and so forth. **Obsolete style files.** The commands for some common packages (such as some used for algorithms), may have changed. Please be certain that you are not compiling your paper using old or obsolete style files.

Final Archive. Place your source files in a single archive which should be compressed using .zip. The final file size may not exceed 10 MB. Name your source file with the last (family) name of the first author, even if that is not you.

Using LaTeX to Format Your Paper

The latest version of the AAAI style file is available on AAAI's website. Download this file and place it in the TeX search path. Placing it in the same directory as the paper should also work. You must download the latest version of the complete AAAI Author Kit so that you will have the latest instruction set and style file.

Document Preamble

In the LaTeX source for your paper, you **must** place the following lines as shown in the example in this subsection. This command set-up is for three authors. Add or subtract author and address lines as necessary, and uncomment the portions that apply to you. In most instances, this is all you need to do to format your paper in the Times font. The helvet package will cause Helvetica to be used for sans serif. These files are part of the PSNFSS2e package, which is freely available from many Internet sites (and is often part of a standard installation).

Leave the setcounter for section number depth commented out and set at 0 unless you want to add section numbers to your paper. If you do add section numbers, you must uncomment this line and change the number to 1 (for section numbers), or 2 (for section and subsection numbers). The style file will not work properly with numbering of subsubsections, so do not use a number higher than 2.

The Following Must Appear in Your Preamble

```
\documentclass[letterpaper]{article}
% DO NOT CHANGE THIS
\usepackage[submission]{aaai24} % DO NOT CHANGE THIS
\usepackage{times} % DO NOT CHANGE THIS
\usepackage{helvet} % DO NOT CHANGE THIS
\usepackage{courier} % DO NOT CHANGE THIS
\usepackage[hyphens]{url} % DO NOT CHANGE THIS
\usepackage{graphicx} % DO NOT CHANGE THIS
\urlstyle{rm} % DO NOT CHANGE THIS
\def\UrlFont{\rm} % DO NOT CHANGE THIS
\usepackage{graphicx}  % DO NOT CHANGE THIS
\usepackage{natbib}  % DO NOT CHANGE THIS
\usepackage{caption}  % DO NOT CHANGE THIS
\frenchspacing % DO NOT CHANGE THIS
\setlength{\pdfpagewidth}{8.5in} % DO NOT CHANGE THIS
\setlength{\pdfpageheight}{11in} % DO NOT CHANGE THIS
%
% Keep the \pdfinfo as shown here. There's no need
% for you to add the /Title and /Author tags.
\pdfinfo{
/TemplateVersion (2024.1)
}
```

Preparing Your Paper

After the preamble above, you should prepare your paper as follows:

```
\begin{document}
\maketitle
\begin{abstract}
%...
\end{abstract}
```

You should then continue with the body of your paper. Your paper must conclude with the references, which should be inserted as follows:

```
% References and End of Paper
% These lines must be placed at the end of your paper
\bibliography{Bibliography-File}
\end{document}
```

```
\begin{document}\\
\maketitle\\
...\\
\bibliography{Bibliography-File}\\
\end{document}\\
```

Commands and Packages That May Not Be Used

There are a number of packages, commands, scripts, and macros that are incompatable with aaai24.sty. The common ones are listed in tables 2 and 3. Generally, if a command, package, script, or macro alters floats, margins, fonts, sizing, linespacing, or the presentation of the references and citations, it is unacceptable. Note that negative vskip and vspace may not be used except in certain rare occurances, and may never be used around tables, figures, captions, sections, subsections, subsubsections, or references.

Page Breaks

For your final camera ready copy, you must not use any page break commands. References must flow directly after the text without breaks. Note that some conferences require references to be on a separate page during the review process. AAAI Press, however, does not require this condition for the final paper.

Paper Size, Margins, and Column Width

Papers must be formatted to print in two-column format on 8.5 x 11 inch US letter-sized paper. The margins must be exactly as follows:

- Top margin: .75 inches
- Left margin: .75 inches
- Right margin: .75 inches
- Bottom margin: 1.25 inches

The default paper size in most installations of LaTeX is A4. However, because we require that your electronic paper be formatted in US letter size, the preamble we have provided includes commands that alter the default to US letter size. Please note that using any other package to alter page size (such as, but not limited to the Geometry package) will result in your final paper being returned to you for correction.

Column Width and Margins. To ensure maximum readability, your paper must include two columns. Each column should be 3.3 inches wide (slightly more than 3.25 inches), with a .375 inch (.952 cm) gutter of white space between the two columns. The aaai24.sty file will automatically create these columns for you.

Overlength Papers

If your paper is too long and you resort to formatting tricks to make it fit, it is quite likely that it will be returned to you. The best way to retain readability if the paper is overlength is to cut text, figures, or tables. There are a few acceptable ways to reduce paper size that don't affect readability. First, turn on \frenchspacing, which will reduce the space after periods. Next, move all your figures and tables to the top of the page. Consider removing less important portions of a figure. If you use \centering instead of \begin{center} in your figure environment, you can also buy some space. For mathematical environments, you may reduce fontsize **but not below 6.5 point**.

Commands that alter page layout are forbidden. These include \columnsep, \float, \topmargin, \topskip, \textheight, \textwidth, \oddsidemargin, and \evensizemargin (this list is not exhaustive). If you alter page layout, you will be required to pay the page fee. Other commands that are questionable and may cause your paper to be rejected include \parindent, and \parskip. Commands that alter the space between sections are forbidden. The title sec package is not allowed. Regardless of the above, if your paper is obviously "squeezed" it is not going to to be accepted. Options for reducing the length of a paper include reducing the size of your graphics, cutting text, or paying the extra page charge (if it is offered).

Type Font and Size

Your paper must be formatted in Times Roman or Nimbus. We will not accept papers formatted using Computer Modern or Palatino or some other font as the text or heading typeface. Sans serif, when used, should be Courier. Use Symbol or Lucida or Computer Modern for *mathematics only.*

Do not use type 3 fonts for any portion of your paper, including graphics. Type 3 bitmapped fonts are designed for fixed resolution printers. Most print at 300 dpi even if the printer resolution is 1200 dpi or higher. They also often cause high resolution imagesetter devices to crash. Consequently, AAAI will not accept electronic files containing obsolete type 3 fonts. Files containing those fonts (even in graphics) will be rejected. (Authors using blackboard symbols must avoid packages that use type 3 fonts.)

Fortunately, there are effective workarounds that will prevent your file from embedding type 3 bitmapped fonts. The easiest workaround is to use the required times, helvet, and courier packages with LaTeX2e. (Note that papers formatted in this way will still use Computer Modern for the mathematics. To make the math look good, you'll either have to use Symbol or Lucida, or you will need to install type 1 Computer Modern fonts — for more on these fonts, see the section "Obtaining Type 1 Computer Modern.")

If you are unsure if your paper contains type 3 fonts, view the PDF in Acrobat Reader. The Properties/Fonts window will display the font name, font type, and encoding properties of all the fonts in the document. If you are unsure if your graphics contain type 3 fonts (and they are PostScript or encapsulated PostScript documents), create PDF versions of them, and consult the properties window in Acrobat Reader.

The default size for your type must be ten-point with twelve-point leading (line spacing). Start all pages (except the first) directly under the top margin. (See the next section for instructions on formatting the title page.) Indent ten points when beginning a new paragraph, unless the paragraph begins directly below a heading or subheading.

Obtaining Type 1 Computer Modern for LaTeX. If you use Computer Modern for the mathematics in your paper (you cannot use it for the text) you may need to download type 1 Computer fonts. They are available without charge from the American Mathematical Society: http://www.ams.org/tex/type1-fonts.html.

Nonroman Fonts. If your paper includes symbols in other languages (such as, but not limited to, Arabic, Chinese, Hebrew, Japanese, Thai, Russian and other Cyrillic languages), you must restrict their use to bit-mapped figures.

Title and Authors

Your title must appear centered over both text columns in sixteen-point bold type (twenty-four point leading). The title must be written in Title Case according to the Chicago Manual of Style rules. The rules are a bit involved, but in general verbs (including short verbs like be, is, using, and go), nouns, adverbs, adjectives, and pronouns should be capitalized, (including both words in hyphenated terms), while articles, conjunctions, and prepositions are lower case unless they directly follow a colon or long dash. You can use the online tool https://titlecaseconverter.com/ to double-check the proper capitalization (select the "Chicago" style and mark the "Show explanations" checkbox).

Author's names should appear below the title of the paper, centered in twelve-point type (with fifteen point leading), along with affiliation(s) and complete address(es) (including electronic mail address if available) in nine-point roman type (the twelve point leading). You should begin the two-column format when you come to the abstract.

Formatting Author Information. Author information has to be set according to the following specification depending if you have one or more than one affiliation. You may not use a table nor may you employ the \authorblk.sty package. For one or several authors from the same institution, please separate them with commas and write all affiliation directly below (one affiliation per line) using the macros \author and \affiliations:

```
\author{
    Author 1, ..., Author n\\
}
\affiliations {
    Address line\\
    ... \\
    Address line\\
}
```

For authors from different institutions, use \textsuperscript {\rm x } to match authors and affiliations. Notice that there should not be any spaces between the author name (or comma following it) and the superscript.

```
\author{
    AuthorOne,\equalcontrib\textsuperscript{\rm 1,\rm2}
    AuthorTwo,\equalcontrib\textsuperscript{\rm 2}
    AuthorThree,\textsuperscript{\rm 3}\\
    AuthorFour,\textsuperscript{\rm 4}
    AuthorFive \textsuperscript{\rm 5}}
}
\affiliations {
    \textsuperscript{\rm 1}AffiliationOne,\\
    \textsuperscript{\rm 2}AffiliationTwo,\\
    \textsuperscript{\rm 3}AffiliationThree,\\
    \textsuperscript{\rm 4}AffiliationFour,\\
```

```
    \textsuperscript{\rm 5}AffiliationFive\\
    \{email, email\}@affiliation.com,
    email@affiliation.com,
    email@affiliation.com,
    email@affiliation.com
}
```

You can indicate that some authors contributed equally using the \equalcontrib command. This will add a marker after the author names and a footnote on the first page.

Note that you may want to break the author list for better visualization. You can achieve this using a simple line break (\\).

LaTeX Copyright Notice

The copyright notice automatically appears if you use aaai24.sty. It has been hardcoded and may not be disabled.

Credits

Any credits to a sponsoring agency should appear in the acknowledgments section, unless the agency requires different placement. If it is necessary to include this information on the front page, use \thanks in either the \author or \title commands. For example:

> \title{Very Important Results in AI\thanks{This work is supported by everybody.}}

Multiple \thanks commands can be given. Each will result in a separate footnote indication in the author or title with the corresponding text at the botton of the first column of the document. Note that the \thanks command is fragile. You will need to use \protect.

Please do not include \pubnote commands in your document.

Abstract

Follow the example commands in this document for creation of your abstract. The command \begin{abstract} will automatically indent the text block. Please do not indent it further. Do not include references in your abstract!

Page Numbers

Do not print any page numbers on your paper. The use of \pagestyle is forbidden.

Text

The main body of the paper must be formatted in black, ten-point Times Roman with twelve-point leading (line spacing). You may not reduce font size or the linespacing. Commands that alter font size or line spacing (including, but not limited to baselinestretch, baselineshift, linespread, and others) are expressly forbidden. In addition, you may not use color in the text.

Citations

Citations within the text should include the author's last name and year, for example (Newell 1980). Append lower-case letters to the year in cases of ambiguity. Multiple authors should be treated as follows: (Feigenbaum and Engelmore 1988) or (Ford, Hayes, and Glymour 1992). In the case of four or more authors, list only the first author, followed by et al. (Ford et al. 1997).

Extracts

Long quotations and extracts should be indented ten points from the left and right margins.

> This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not necessary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

Footnotes

Use footnotes judiciously, taking into account that they interrupt the reading of the text. When required, they should be consecutively numbered throughout with superscript Arabic numbers. Footnotes should appear at the bottom of the page, separated from the text by a blank line space and a thin, half-point rule.

Headings and Sections

When necessary, headings should be used to separate major sections of your paper. Remember, you are writing a short paper, not a lengthy book! An overabundance of headings will tend to make your paper look more like an outline than a paper. The aaai24.sty package will create headings for you. Do not alter their size nor their spacing above or below.

Section Numbers. The use of section numbers in AAAI Press papers is optional. To use section numbers in LaTeX, uncomment the setcounter line in your document preamble and change the 0 to a 1. Section numbers should not be used in short poster papers and/or extended abstracts.

Section Headings. Sections should be arranged and headed as follows:

1. Main content sections
2. Appendices (optional)
3. Ethical Statement (optional, unnumbered)
4. Acknowledgements (optional, unnumbered)
5. References (unnumbered)

Appendices. Any appendices must appear after the main content. If your main sections are numbered, appendix sections must use letters instead of arabic numerals. In LaTeX you can use the `\appendix` command to achieve this effect and then use `\section{Heading}` normally for your appendix sections.

Ethical Statement. You can write a statement about the potential ethical impact of your work, including its broad societal implications, both positive and negative. If included, such statement must be written in an unnumbered section titled *Ethical Statement*.

Acknowledgments. The acknowledgments section, if included, appears right before the references and is headed "Acknowledgments". It must not be numbered even if other sections are (use `\section*{Acknowledgements}` in LaTeX). This section includes acknowledgments of help from

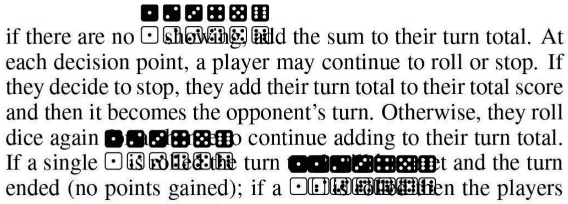

Figure 1: Using the trim and clip commands produces fragile layers that can result in disasters (like this one from an actual paper) when the color space is corrected or the PDF combined with others for the final proceedings. Crop your figures properly in a graphics program – not in LaTeX.

associates and colleagues, credits to sponsoring agencies, financial support, and permission to publish. Please acknowledge other contributors, grant support, and so forth, in this section. Do not put acknowledgments in a footnote on the first page. If your grant agency requires acknowledgment of the grant on page 1, limit the footnote to the required statement, and put the remaining acknowledgments at the back. Please try to limit acknowledgments to no more than three sentences.

References. The references section should be labeled "References" and must appear at the very end of the paper (don't end the paper with references, and then put a figure by itself on the last page). A sample list of references is given later on in these instructions. Please use a consistent format for references. Poorly prepared or sloppy references reflect badly on the quality of your paper and your research. Please prepare complete and accurate citations.

Illustrations and Figures

Your paper must compile in PDFLaTeX. Consequently, all your figures must be .jpg, .png, or .pdf. You may not use the .gif (the resolution is too low), .ps, or .eps file format for your figures.

Figures, drawings, tables, and photographs should be placed throughout the paper on the page (or the subsequent page) where they are first discussed. Do not group them together at the end of the paper. If placed at the top of the paper, illustrations may run across both columns. Figures must not invade the top, bottom, or side margin areas. Figures must be inserted using the \usepackage{graphicx}. Number figures sequentially, for example, figure 1, and so on. Do not use minipage to group figures.

If you normally create your figures using pgfplots, please create the figures first, and then import them as pdfs with proper bounding boxes, as the bounding and trim boxes created by pfgplots are fragile and not valid.

When you include your figures, you must crop them **outside** of LaTeX. The command \includegraphics*[clip=true, viewport 0 0 10 10]... might result in a PDF that looks great, but the image is **not really cropped.** The full image can reappear (and obscure whatever it is overlapping) when page numbers are applied or color space is standardized. Figures 1, and 2 display some unwanted results that often occur.

If your paper includes illustrations that are not compatible with PDFTEX (such as .eps or .ps documents), you will need to convert them. The epstopdf package will usually work for eps files. You will need to convert your ps files to PDF in either case.

Figure Captions. The illustration number and caption must appear *under* the illustration. Labels and other text with the actual illustration must be at least nine-point type. However, the font and size of figure captions must be 10 point roman. Do not make them smaller, bold, or italic. (Individual words may be italicized if the context requires differentiation.)

Tables

Tables should be presented in 10 point roman type. If necessary, they may be altered to 9 point type. You may not use any commands that further reduce point size below nine points. Tables that do not fit in a single column must be placed across double columns. If your table won't fit within the margins even when spanning both columns, you must split it. Do not use minipage to group tables.

Table Captions. The number and caption for your table must appear *under* (not above) the table. Additionally, the font and size of table captions must be 10 point roman and must be placed beneath the figure. Do not make them smaller, bold, or italic. (Individual words may be italicized if the context requires differentiation.)

Low-Resolution Bitmaps. You may not use low-resolution (such as 72 dpi) screen-dumps and GIF files—these files contain so few pixels that they are always blurry, and illegible when printed. If they are color, they will become an indecipherable mess when converted to black and white. This is always the case with gif files, which should never be used. The resolution of screen dumps can be increased by reducing the print size of the original file while retaining the same number of pixels. You can also enlarge files by manipulating them in software such as PhotoShop. Your figures should be 300 dpi when incorporated into your document.

LaTeX Overflow. LaTeX users please beware: LaTeX will sometimes put portions of the figure or table or an equation in the margin. If this happens, you need to make the figure or table span both columns. If absolutely necessary, you may reduce the figure, or reformat the equation, or reconfigure the table. **Check your log file!** You must fix any overflow into the margin (that means no overfull boxes in LaTeX). **Nothing is permitted to intrude into the margin or gutter.**

Using Color. Use of color is restricted to figures only. It must be WACG 2.0 compliant. (That is, the contrast ratio must be greater than 4.5:1 no matter the font size.) It must be CMYK, NOT RGB. It may never be used for any portion of the text of your paper. The archival version of your paper will be printed in black and white and grayscale. The web version must be readable by persons with disabilities. Consequently, because conversion to grayscale can cause unde-

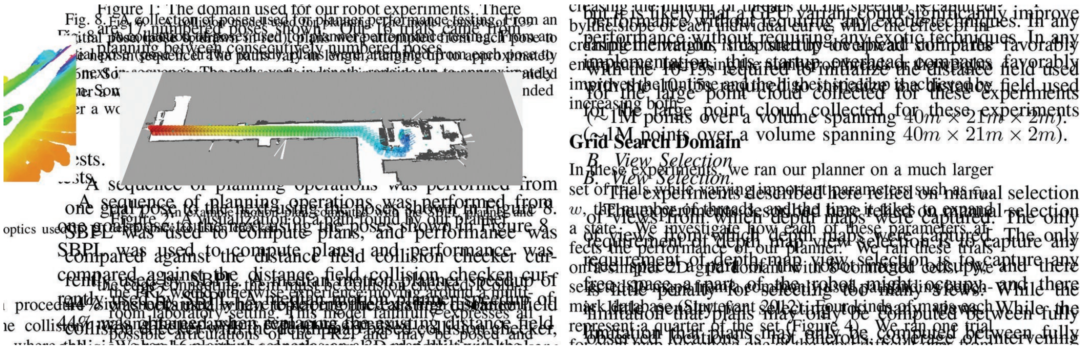

Figure 2: Adjusting the bounding box instead of actually removing the unwanted data resulted multiple layers in this paper. It also needlessly increased the PDF size. In this case, the size of the unwanted layer doubled the paper's size, and produced the following surprising results in final production. Crop your figures properly in a graphics program. Don't just alter the bounding box.

sirable effects (red changes to black, yellow can disappear, and so forth), we strongly suggest you avoid placing color figures in your document. If you do include color figures, you must (1) use the CMYK (not RGB) colorspace and (2) be mindful of readers who may happen to have trouble distinguishing colors. Your paper must be decipherable without using color for distinction.

Drawings. We suggest you use computer drawing software (such as Adobe Illustrator or, (if unavoidable), the drawing tools in Microsoft Word) to create your illustrations. Do not use Microsoft Publisher. These illustrations will look best if all line widths are uniform (half- to two-point in size), and you do not create labels over shaded areas. Shading should be 133 lines per inch if possible. Use Times Roman or Helvetica for all figure call-outs. **Do not use hairline width lines** — be sure that the stroke width of all lines is at least .5 pt. Zero point lines will print on a laser printer, but will completely disappear on the high-resolution devices used by our printers.

Photographs and Images. Photographs and other images should be in grayscale (color photographs will not reproduce well; for example, red tones will reproduce as black, yellow may turn to white, and so forth) and set to a minimum of 300 dpi. Do not prescreen images.

Resizing Graphics. Resize your graphics **before** you include them with LaTeX. You may **not** use trim or clip options as part of your \includegraphics command. Resize the media box of your PDF using a graphics program instead.

Fonts in Your Illustrations. You must embed all fonts in your graphics before including them in your LaTeX document.

Algorithms. Algorithms and/or programs are a special kind of figures. Like all illustrations, they should appear floated to the top (preferably) or bottom of the page. However, their caption should appear in the header, left-justified and enclosed between horizontal lines, as shown in Algorithm 1. The algorithm body should be terminated with another horizontal line. It is up to the authors to decide whether

Algorithm 1: Clustering algorithm

Input: Columns selected for clustering
Output: Your algorithm's output

```
Place set of N instances in a single cluster k.
Calculate the overall error rate M_k
for i = 1 until max iterations do
  Do some action.
  if conditional then
    Perform task A.
  else
    Perform task B.
  end if
end while
return solution
```

Listing 1: Example listing `quicksort.hs`

```
quicksort :: Ord a => [a] -> [a]
quicksort []     = []
quicksort (p:xs) = (quicksort lesser) ++
      [p] ++ (quicksort greater)
  where
    lesser  = filter (< p) xs
    greater = filter (>= p) xs
```

to show line numbers or not, how to format comments, etc.

In LaTeX algorithms may be typeset using the `algorithm` and `algorithmic` packages, but you can also use one of the many other packages for the task.

Listings. Listings are much like algorithms and programs. They should also appear floated to the top (preferably) or bottom of the page. Listing captions should appear in the header, left-justified and enclosed between horizontal lines as shown in Listing 1. Terminate the body with another horizontal line and avoid any background color. Line numbers, if included, must appear within the text column.

References

The AAAI style includes a set of definitions for use in formatting references with BibTeX. These definitions make the bibliography style fairly close to the ones specified in the Reference Examples appendix below. To use these definitions, you also need the BibTeX style file "aaai24.bst," available in the AAAI Author Kit on the AAAI web site. Then, at the end of your paper but before \enddocument, you need to put the following lines:

\bibliography{bibfile1,bibfile2,...}

Proofreading Your PDF

Please check all the pages of your PDF file. The most commonly forgotten element is the acknowledgements — especially the correct grant number. Authors also commonly forget to add the metadata to the source, use the wrong reference style file, or don't follow the capitalization rules or comma placement for their author-title information properly. A final common problem is text (expecially equations) that runs into the margin. You will need to fix these common errors before submitting your file.

Improperly Formatted Files

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb). Consequently, if your file is improperly formatted, it will be returned to you for correction.

Naming Your Electronic File

We require that you name your LaTeX source file with the last name (family name) of the first author so that it can easily be differentiated from other submissions. Complete file-naming instructions will be provided to you in the submission instructions.

Submitting Your Electronic Files to AAAI

Instructions on paper submittal will be provided to you in your acceptance letter.

Inquiries

If you have any questions about the preparation or submission of your paper as instructed in this document, please contact AAAI Press at the address given below. If you have technical questions about implementation of the aaai style file, please contact an expert at your site. We do not provide technical support for LaTeX or any other software package. To avoid problems, please keep your paper simple, and do not incorporate complicated macros and style files.

AAAI Press
1900 Embarcadero Road, Suite 101
Palo Alto, California 94303-3310 USA
Telephone: (650) 328-3123
E-mail: See the submission instructions for your particular conference or event.

Additional Resources

LaTeX is a difficult program to master. If you've used that software, and this document didn't help or some items were not explained clearly, we recommend you read Michael Shell's excellent document (testflow doc.txt V1.0a 2002/08/13) about obtaining correct PS/PDF output on LaTeX systems. (It was written for another purpose, but it has general application as well). It is available at www.ctan.org in the tex-archive.

Reference Examples

* Formatted bibliographies should look like the following examples. You should use BibTeX to generate the references. Missing fields are unacceptable when compiling references, and usually indicate that you are using the wrong type of entry (BibTeX class).

Book with multiple authors Use the `@book` class.

em:86.

Journal and magazine articles Use the `@article` class.

r:80.

hcr:83.

Proceedings paper published by a society, press or publisher Use the `@inproceedings` class. You may abbreviate the *booktitle* field, but make sure that the conference edition is clear.

c:84.

c:83.

University technical report Use the `@techreport` class.

r:86.

Dissertation or thesis Use the `@phdthesis` class.

c:79.

Forthcoming publication Use the `@misc` class with a `note="Forthcoming"` annotation.

```
@misc(key,
  [...]
  note="Forthcoming",
)
```

c:21.

ArXiv paper Fetch the BibTeX entry from the "Export Bibtex Citation" link in the arXiv website. Notice it uses the `@misc` class instead of the `@article` one, and that it includes the `eprint` and `archivePrefix` keys.

```
@misc(key,
  [...]
  eprint="xxxx.yyyy",
  archivePrefix="arXiv",
)
```

c:22.

Website or online resource Use the `@misc` class. Add the url in the `howpublished` field and the date of access in the `note` field:

```
@misc(key,
  [...]
  howpublished="\url{http://...}",
  note="Accessed: YYYY-mm-dd",
)
```

c:23.

For the most up to date version of the AAAI reference style, please consult the *AI Magazine* Author Guidelines at https://aaai.org/ojs/index.php/aimagazine/about/submissions#authorGuidelines

Acknowledgments

Civic AI lab

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 648–657.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Chakraborty, J.; Peng, K.; and Menzies, T. 2020. Making fair ML software using trustworthy explanation. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, 1229–1233.

Dimanov, B.; Bhatt, U.; Jamnik, M.; and Weller, A. 2020. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*, 2473–2480. IOS Press.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Helwegen, R.; Louizos, C.; and Forré, P. 2020. Improving fair predictions using variational inference in causal models. *arXiv preprint arXiv:2008.10880*.

Huang, Z. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. *Dmkd*, 3(8): 34–39.

Jain, A.; Ravula, M.; and Ghosh, J. 2020. Biased models have biased explanations. *arXiv preprint arXiv:2012.10986*.

Ji, J.; Pang, W.; Zhou, C.; Han, X.; and Wang, Z. 2012. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 30: 129–135.

Kumar, I. E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, 5491–5500. PMLR.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Luong, B. T.; Ruggieri, S.; and Turini, F. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Misztal-Radecka, J.; and Indurkhya, B. 2021. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management*, 58(3): 102519.

Molnar, C. 2023. *Interpreting Machine Learning Models With SHAP: A Guide With Python Examples And Theory On Shapley Values*.

Muhammad, S. 2021. *Algorithmic Fairness with Unsupervised Bias Discovery*. Master's thesis, Vrije Universiteit Amsterdam.

Nasiriani, N.; Squicciarini, A.; Saldanha, Z.; Goel, S.; and Zannone, N. 2019. Hierarchical clustering for discrimination discovery: A top-down approach. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 187–194. IEEE.

Silberg, J.; and Manyika, J. 2019. Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute*, 1(6): 1–31.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Case description	Case label	SHAP (SH)	Regular (R)	Error (E)	Sensitive (SE)	K-means (K)	K-Prototypes (P)	DBSCAN (D)
Baseline	B_K		X		X	X		
Baseline (+E)	BE_K		X	X	X	X		
Baseline	B_P		X		X		X	
Baseline (+E)	BE_P		X	X	X		X	
Baseline	B_D		X		X			X
Baseline (+E)	BE_D		X	X	X			X
SHAP enhanced	SHB_K	X	X		X	X		
SHAP enhanced (+E)	SHBE_K	X	X	X	X	X		
SHAP enhanced	SHB_P	X	X		X		X	
SHAP enhanced (+E)	SHBE_P	X	X	X	X		X	
SHAP enhanced	SHB_D	X	X		X			X
SHAP enhanced (+E)	SHBE_D	X	X	X	X			X
SHAP-only	SH_K	X				X		
SHAP-only (+E)	SHE_K	X		X		X		
SHAP-only	SH_P	X					X	
SHAP-only(+E)	SHE_P	X		X			X	
SHAP-only	SH_D	X						X
SHAP-only (+E)	SHE_D	X		X				X
SENS only	SE_K				X	X		
SENS (+E)	SEE_K			X	X	X		
SENS (+SHAP)	SESH_K	X			X	X		
SENS(+E+SHAP)	SEESH_K	X		X	X	X		
SENS only	SE_P				X		X	
SENS (+E)	SEE_P			X	X		X	
SENS (+SHAP)	SESH_P	X			X		X	
SENS(+E+SHAP)	SEESH_P	X		X	X		X	
SENS only	SE_D				X			X
SENS (+E)	SEE_D			X	X			X
SENS (+SHAP)	SESH_D	X			X			X
SENS(+E+SHAP)	SEESH_D	X		X	X			X

Table 1: Feature Combinations per Clustering algorithm

\abovecaption	\abovedisplay	\addevensidemargin	\addsidemargin
\addtolength	\baselinestretch	\belowcaption	\belowdisplay
\break	\clearpage	\clip	\columnsep
\float	\input	\input	\linespread
\newpage	\pagebreak	\renewcommand	\setlength
\text height	\tiny	\top margin	\trim
\vskip{-	\vspace{-		

Table 2: Commands that must not be used

authblk	babel	cjk	dvips
epsf	epsfig	euler	float
fullpage	geometry	graphics	hyperref
layout	linespread	lmodern	maltepaper
navigator	pdfcomment	pgfplots	psfig
pstricks	t1enc	titlesec	tocbind
ulem			

Table 3: LaTeX style packages that must not be used.