# Wanted:
# Personalised Bias Warnings for Gender Bias in Language Models

Chiara Di Bonaventura[1,2], Michelle Nwachukwu[1,2], Maria Stoica[1,2]

[1]King's College London, [2]Imperial College London
{chiara.di_bonaventura, michelle.nwachukwu}@kcl.ac.uk, m.stoica22@imperial.ac.uk

## Abstract

The widespread use of language models, especially Large Language Models, paired with their inherent biases can propagate and amplify societal inequalities. While research has extensively explored methods for bias mitigation and measurement, limited attention has been paid to how such biases are communicated to users, which instead can have a positive impact on increasing user trust and understanding of these models. Our study addresses this gap by investigating user preferences for gender bias mitigation, measurement and communication in language models. To this end, we conducted a user study targeting female AI practitioners with eighteen female and one male participant. Our findings reveal that user preferences for bias mitigation and measurement show strong consensus, whereas they vary widely for bias communication, underscoring the importance of tailoring warnings to individual needs. Building on these findings, we propose a framework for user-centred bias reporting, which leverages runtime monitoring techniques to assess and visualise bias in real time and in a customizable fashion.

## 1 Introduction

Many practitioners use Large Language Models (LLMs) in everyday applications, like conversational agents, due to their accessibility. They are primarily hosted in large infrastructures such as Hugging Face[1] and require a few lines of code. However, their wide adoption comes with some limitations and risks which might be overlooked or not entirely understood by practitioners (Bianchi and Hovy, 2021; Weidinger et al., 2022; Bianchi et al., 2023a).

In this context, socio-demographic bias in language models is a well-known issue which has gained much attention following the paradigm shift in the development of language models from a performance-based to a transparency-based perspective (Sap et al., 2020; Blacklaws, 2018). In particular, gender bias is the most investigated type of sociodemographic bias (Gupta et al., 2024). Most of the research in Natural Language Processing (NLP) focuses either on bias mitigation or bias detection (Blodgett et al., 2020). The former has proposed several techniques to de-bias language models (e.g., Mahabadi et al. (2020); Utama et al. (2020)). The latter instead has led to the development of many resources like datasets and tests to analyse whether and to what extent language models are biased (e.g., Nadeem et al. (2021); Caliskan et al. (2017)). Practitioners can use these resources to understand the limitations and risks behind LLMs, which should ideally guide their decision when choosing an LLM to adopt. However, the current literature lacks a user-centred approach to bias in language models.

While few studies have suggested frameworks to publicly inform practitioners about the presence of bias within a language model (Nozza et al., 2022) or assess the actionability of a certain bias measure (Delobelle et al., 2024), the user perspective around bias in NLP is often neglected. This is a central aspect to consider when developing resources to either detect or mitigate bias in language models, as it can increase not only the practitioners' understanding of language models' limitations but also their trust in these models (Gaba et al., 2023). Therefore, in this work, we seek to understand practitioners' perspectives regarding *(i)* bias mitigation (i.e., when to intervene to reduce bias), *(ii)* bias measurement (i.e., which metrics to use to measure bias), and *(iii)* bias warnings (i.e., how to inform about the presence of bias) in the context of language models.

**Contributions.** Our contribution is twofold. **(1)** We conduct a user study targeting female practition-

[1] https://huggingface.co/

*Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 124–136
August 1, 2025 ©2025 Association for Computational Linguistics

ers during a workshop promoting gender-inclusive AI systems to collect their perspectives on socio-demographic biases in language models, focusing especially on gender bias. **(2)** We propose a customizable framework to monitor bias in language models grounded on the findings of our study.

## 2 Bias Statement

We focus on socio-demographic biases, particularly gender bias, where we consider system behaviours to be biased when they systematically produce skewed or unfair results like, for instance, reproducing or amplifying harmful stereotypes, erasing marginalised identities, or unequally treating female and male groups. These behaviours are harmful because they can reinforce existing social inequalities, especially if we consider the widespread adoption of language models by practitioners across many domains. In Section 6, we discuss an example in the financial sector but similar implications can hold in other sectors as well.

## 3 Related Work

Following, we discuss existing research on socio-demographic biases in NLP research, 'bias warnings' and user-centred studies in the field.

**Socio-demographic Biases in NLP research.** Research on bias in language models is an active field in NLP research, with most of the work focusing on socio-demographic biases (Lauscher et al. (2022a); Hung et al. (2023); Cercas Curry et al. (2024), *inter alia*). According to a recent survey of Gupta et al. (2024), gender bias is the most investigated type of socio-demographic biases among other types, like race, ethnicity, or age. Research in this field has led to several studies investigating whether and to what extent language models are biased (i.e., ***bias measurement***). Examples include machine translation (e.g., Bianchi et al. (2023b)), text classification (e.g., Sobhani and Delany (2024)), speech recognition (e.g., Attanasio et al. (2024)), visual question answering (e.g., Ruggeri and Nozza (2023)). These studies adopt either extrinsic or intrinsic metrics to quantify how biased language models are. The former look at the representational level inside the model (e.g., Word Embeddings Association Test (WEAT) (Caliskan et al., 2017)), whereas the latter focus on the behavioural level in downstream tasks (e.g., subgroup Area-Under-the-Curve (AUC) (Borkan et al., 2019)). In addition to measuring bias, several NLP studies have proposed de-biasing techniques to reduce bias within language models (i.e., ***bias mitigation***). The de-biasing approaches can be broadly categorised as data-centric and model-centric approaches. The former are techniques that manipulate the input data before running a standard model training procedure (Le Bras et al. (2020); Min et al. (2020), *inter alia*). The latter are de-biasing techniques that either modify the architecture of the model, the optimisation, or the training procedure in order to reduce the model's reliance on spurious biases (Sagawa et al. (2019); Tu et al. (2020), *inter alia*). Despite all these efforts to comprehensively measure and mitigate bias in language models, we currently lack an understanding of how practitioners perceive bias. This work addresses this gap by conducting a user study on gender bias in language models, targeting female practitioners. Additionally, we investigate whether their perspectives change based on the type of bias, i.e., gender bias vs. other socio-demographic biases.

**Bias Warnings.** While bias measurement and bias mitigation are widely investigated in NLP research (Blodgett et al., 2020), fewer studies have focused on how to warn practitioners about the presence of bias within language models (i.e., ***bias warning***). We group all the resources proposed to inform practitioners under the term 'bias warnings.' Several studies have proposed attaching additional information to datasets, explaining data characteristics, limitations, and best use cases. Examples include data cards (Pushkarna et al., 2022), datasheets (Gebru et al., 2021), and meta-data formats like Croissant (Akhtar et al., 2025). Similarly, some studies have proposed model cards that detail how the model is trained, evaluated, and intended to be used (Mitchell et al., 2019). Instead of adding documentation, recent studies have proposed frameworks to actively inform practitioners. Nozza et al. (2022) suggest social bias tests in model development pipelines to verify how biased and harmful language models are. According to this framework, models should be released with a badge system that identifies possible issues that practitioners might encounter with the model. Delobelle et al. (2024) propose a framework of desiderata for actionability in bias measures, i.e., what information is required of a bias measure to enable practitioners to act based on its results. However, studies on bias warnings adopt a one-size-fits-all strategy, which may not meet the diverse user expectations and needs. For instance, a technologically savvy user might prefer a different bias warning than a non-expert user. In this work, we first assess individual preferences about bias and then develop a personalised framework for bias warnings.

**User-Centred Studies.** Recent studies have investigated the impact of specific bias warnings on user trust and decision-making in a wide set of AI systems, from recommendation systems (Doppalapudi et al., 2024) to standard machine learning models (Gaba et al., 2023; Cabrera et al., 2023). Others have focused on data and model documentation. For instance, Crisan et al. (2022) expanded the traditionally static model cards by suggesting an interactive framework where practitioners can, for example, observe data distribution or play with examples in real time. Their interactive framework is shown to benefit users, especially those who are non-experts. Focusing on language models instead, most of the proposed bias warnings are not tested on users, which limits their potential impact. Indeed, recent research on individual user preferences in LLMs shows a misalignment between expected and contextual preferences (Kirk et al., 2024; Di Bonaventura et al., 2024), where expected preferences are those stated by users before engaging with the model, whereas contextual preferences are those stated by the users after having engaged with the model. We fill this gap by proposing a user-centred study on socio-demographic biases in language models; these findings are used to present a personalised monitoring framework for bias warnings.

## 4 User Study

In June 2024, we conducted a pilot study at an ACM WomENcourage[2] workshop that aimed to promote gender-inclusive AI systems by fostering interdisciplinary dialogue and ethical reflection. ACM WomENcourage is an event that celebrates the contributions of women in computing and supports professionals at different stages of their careers. In 2024, the theme of the event was Responsible Computing for Gender Equality, highlighting the gender gap in technology and advocating for computing tools for social progress. Our workshop was structured to address the critical intersection of gender bias and language models. Through a combination of theoretical presentations, hands-on activities, and discussions, participants were introduced to how to identify, measure, and mitigate gender bias in language models. Specifically, the workshop presentation was split into two parts: Bias Mitigation (Section 4.1) and Bias Measurement (Section 4.2), followed by the Pilot Study (Section 4.3).

### 4.1 Bias Mitigation: How does gender bias enter language models' pipelines?

Bias in AI systems like language models can appear at different stages of the system's development pipeline (Hovy and Prabhumoye, 2021; Gallegos et al., 2024), including data collection, model development, and evaluation.

**1. Data Collection.** Training data often reflects existing social imbalances. For example, if one group is overrepresented in the data, the system may unfairly favour that group. Similarly, underrepresentation can lead to poor performance for minority groups (Mehrabi et al., 2021). For instance, in Wikipedia, which has widely been used to train language models, only 15.5% of English bios are about women (Wagner et al., 2016). In addition to imbalanced data, there is the issue of stereotypical representation: even when minorities are present in the data, they are often represented stereotypically and/or suffer from biased sampling. For example, queer and lesbian people are more often associated with toxic comments than neutral comments (Dixon et al., 2018).

**2. Model Development.** During training, language models learn biased word representations not only from the imbalanced, stereotypical and biased representations in the datasets but also from the decisions made during system development, which can amplify biases (Ziosi et al., 2024; Buda et al., 2024; Nino and Lisi, 2024). Examples include optimising solely for accuracy without considering fairness (Rueda et al., 2024). This results in language models, for instance, translating "He is a nurse. She is a doctor." to Hungarian and back to English as "She is a nurse. He is a doctor." (Douglas, 2017). Or, in language models trained for sentiment analysis, texts mentioning female terms are more likely to be associated with anger than those containing male terms (Park et al., 2018). Similarly, in story generation, language models are shown to complete a story in which the male protagonist earned a college degree while the female

[2] https://womencourage.acm.org/2024/

protagonist made spaghetti (Huang et al., 2021).

**3. Evaluation.** Bias in language models extends beyond data and model behavior to the evaluation stage itself, as testing processes, annotation guidelines, and annotator demographics can introduce or reinforce biased outcomes. Testing processes may not account for the full range of biases, particularly when fairness is measured in overly simplistic ways, such as focusing on binary categories and ignoring intersectional factors like race and gender combined (Tyser et al., 2024). Moreover, the groundtruth used to evaluate models often reflects the dominant perspective, failing to account for the subjective viewpoints of different socio-demographic groups (Orlikowski et al., 2025). Examples include the fact that belonging to LGBTQ identities impacts annotators' behaviours concerning homophobic content (Goyal et al., 2022).

Throughout this 3-step pipeline, several challenges can hinder the mitigation of bias, making this a complex issue to handle. Binary thinking is a challenge that distils fairness into a comparison between two groups. This oversimplifies the experiences of people from identities that fall beyond the binary (Barocas et al., 2023). This also does not consider intersectionality, so binary thinking can ignore those affected by both racial and gender bias (Buolamwini and Gebru, 2018). Another complex challenge is how to define harms. The focus is often placed on unequal outcomes, but reinforcement of stereotypes and lack of representation for particular groups can also be harmful (Mehrabi et al., 2021). Mitigating bias in AI requires a careful balance between technical solutions and a broader understanding of societal inequalities.

### 4.2 Bias Measurement: How do we identify gender bias in language models?

Currently, two paradigms exist to measure bias: intrinsic and extrinsic (Gallegos et al., 2024; Li et al., 2023). The former examines the representational level inside the model, whereas the latter examines the behavioural level in downstream tasks.

**1. Intrinsic Metrics.** Clustering techniques are widely used to understand how the model represents concepts and identify potentially biased patterns. For example, Gonen and Goldberg (2019) measures gender bias in language models using cluster bias of a target word *w*, which is calculated as the percentage of male and female stereotypical words among the k nearest neighbours of *w*'s embedding. Word Embeddings Association Test (WEAT) (Caliskan et al., 2017) is another established intrinsic bias measure, which quantifies bias using semantic similarities between word embeddings across ten bias tests. Each test specifies two sets of target words $t$ (e.g., male and female words), and two sets of attributes $a$ (e.g., career- and family-related words). The bias is then measured as the difference in the association strength between $t_1, a_1$ and $t_1, a_2$ and with respect to their $t_2$ counterparts. Another intrinsic measure is ad hoc probes designed to identify how much the model representations align with potentially harmful patterns, like stereotypes. Examples include StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) tests, where the model is asked to fill-in the blank space in testing sentences, and it is then evaluated on its tendency to generate stereotypical or anti-stereotypical sentences.

**2. Extrinsic Metrics.** Most of them focus on group-specific performance, quantifying group disparity in downstream tasks: subgroup Area-Under-the-Curve (AUC) (Borkan et al., 2019), False Positive, False Negative Equality Difference (Dixon et al., 2018), Predictive Parity, Equal Opportunity Difference. Recently, some studies have adopted explainable methods to measure bias in downstream tasks. For instance, Attanasio et al. (2022) uses post-hoc token-level explanations to explain which words in the input text were responsible for the model prediction, highlighting how Transformer-based models (Vaswani, 2017) often misclassify neutral texts as misogynous texts due to their overreliance on biased keywords. In this case, models' bias is measured using plausibility and faithfulness metrics (Jacovi and Goldberg, 2020), which evaluate how much the explanations are aligned with human beliefs and model reasoning, respectively.

While it should be desirable for a system to have low intrinsic and extrinsic bias metrics, this is often not the case. Indeed, recent work has shown how fixing one metric does not necessarily resolve the other, as they are not positively correlated (Goldfarb-Tarrant et al., 2021). Therefore, the choice between which metrics to prioritise is left to a trade-off: task-free but not easily quantifiable intrinsic metrics or easily quantifiable but task-constrained extrinsic metrics?

### 4.3 Pilot Study

Following the presentation on bias mitigation and bias measurement, we conducted a pilot study to discuss and collect feedback on bias mitigation, measurement, and warnings from AI practitioners, specifically targeting female practitioners. We recruited participants for the pilot study from attendees of our workshop. We introduced the study at the beginning of the workshop to give attendees time to decide if they wished to participate. At the end of the workshop presentations, those that were interested in taking part were given more information and signed a consent form before their participation. This study was approved by the main authors' institution's College Research Ethics Committee (CREC).

**Participants.** Nineteen participants took part in our pilot study, including eighteen women and one man. The overwhelming participation of women was expected as the workshop was held at a conference specifically aimed at celebrating the role and impact of women in computing. We note that our study focuses on binary gender categories, reflecting the demographic composition of the workshop attendees. As such, it does not capture perspectives from non-binary or transgender individuals, which we acknowledge as a limitation and an important direction for future work (Lauscher et al., 2022b, 2023). Participants had varying levels of expertise with language models. Most participants self-identified as advanced beginners, with five considering themselves novices and eight as advanced beginners. Three participants rated themselves as competent, and another three as proficient, while none identified as experts.

**Pilot Study Overview.** The pilot study sought to evaluate the workshop's effectiveness and gain insights into participants' perceptions of gender bias in AI systems. Three questions were posed to 19 participants, encouraging critical reflection on bias intervention, measurement, and communication. Participants were asked to fill out a form asking about their level of expertise in language models, their gender identity, and the following open-ended questions (Q).

**Q1:** *Considering the whole pipeline to create a system like a language model (i.e., data curation, development, and evaluation), which step is the most important to intervene in to reduce gender bias? Do you think your answer would be different depending on the type of bias? Why?*

This question aimed to identify critical stages in the language models' pipeline where interventions would have the greatest impact on reducing gender bias. At the same time, we wanted to assess whether practitioners' choices would change based on the type of socio-demographic bias.

**Q2:** *Considering intrinsic and extrinsic metrics, which do you believe is more effective for measuring gender bias in language models? Should we look 'inside' these models (i.e., intrinsic) or should we look at how these models 'behave' in a downstream application (i.e., extrinsic)? Do we need both? If yes, why? If not, which is best?*

Participants were prompted to evaluate the effectiveness of intrinsic and extrinsic measures for detecting gender bias and consider the necessity of using both approaches.

**Q3:** *How would you like to be informed about the presence of gender bias in a language model? Examples might include reporting the score on a standardized external benchmark, the number of tests successfully passed in a series of safety tests, visualizing biased examples within the system, other...*

This question was designed to explore individual preferences for reporting of gender bias in language models to effectively inform practitioners.

## 5 Findings

In the following sections, we discuss the main findings of our pilot study, grouped by question.

### 5.1 Q1: Bias Mitigation

All participants considered data curation the most important step to intervene in the language models' pipeline to mitigate gender bias (Table 1), ensuring that all groups get a *fair* representation in the data (i.e., balanced, non-stereotypical, and as unbiased as possible). Indeed, LLMs are particularly susceptible to such biases, as they rely heavily on the data they are trained on. Participants seemed to have a strong understanding of how input data can affect the performance of language models. As one participant put it, "CICO (Crap In, Crap Out) underscores the importance of careful dataset curation to mitigate bias.". Moreover, participants emphasised that mitigating bias is hard to define, as what is considered bias is often context-dependent. Some noted that cultural and historical patterns are often reflected in data, and biases present can be passed on to the models, affecting their output. One participant pointed out that while associations like ‘female’ with ‘home worker’ and ‘male’ with ‘career’ may reflect historical realities that are not appropriate for today, the presence of these historical associations may be helpful depending on the application.

Four people also mentioned the evaluation stage of the language models’ pipeline as an important bias mitigation step. One participant pointed that evaluating language models with fairness metrics in addition to standard performance metrics and/or accounting for subjectivity can potentially catch what was missed during data curation, “this way one can iterate on the development of the model and keep improving it.”. Similarly, another participant said “I give more weight to data curation kind of as a filter and then evaluation to refine the model.”.

Lastly, participants were asked if their choice would change based on the type of bias, i.e., gender vs. other socio-demographic biases. Most of the participants said that the type of bias would not affect their answers. However, they acknowledged that their answers could differ depending on the use case of language models. For instance, one participant reported that in the medical domain, mitigating bias during model development (e.g., using fairness optimisation) is better than data curation. Others have focused on machine translation and gendered vs. non-gendered languages, reporting that “datasets should be altered for an inclusive language” (i.e., data curation) for gendered languages like German and Spanish whereas for non-gendered languages “the best way to tell if there is discriminatory outcomes is in the evaluation stage, potentially going back to mitigate in the development stage.”. Participants’ attention to the application of language models rather than their type of socio-demographic bias aligns with previous studies showing how socio-demographic attributes matter based on the context rather than the type of socio-demographic itself (Gaci, 2023). Indeed, there are high-stake scenarios like medical and legal where mitigating for socio-demographic biases is crucial—the so-called *undesired* subjectivity—whereas other domains like conversational agents where some degree of socio-demographic tailoring is considered appropriate or even desirable—the so-called *desired* subjectivity.

| | Number Selected |
|---|---|
| Data curation | 19 |
| Evaluation | 4 |
| Development | 1 |

Table 1: Results from the pilot study for bias mitigation. Note that we allowed participants to choose multiple answers.

### 5.2 Q2: Bias Measurement

The majority of participants said that both intrinsic and extrinsic metrics serve distinct but valuable purposes, with 63% stating this as their preference. In this case, a few participants distinguished between the individual contributions of the two measures: intrinsic measures are often used by researchers and engineers to understand model behaviour and refine performance, while extrinsic evaluations are critical for assessing broader societal impacts. Some highlighted that extrinsic measures are more important for determining specific user outcomes, but intrinsic evaluations provide valuable insights into the overall behaviour of a language model. Others noted that different aspects of bias are measured by each method, making a combined approach necessary for a more comprehensive understanding of the bias of a given model. Additionally, one participant suggested that justice theories from philosophy should inform both model design and evaluation processes. One of the participants commented that: “We need both, but for different uses. Intrinsic measures can help give insights to systems or their use. Extrinsic measures are overall more crucial because they are the ones that capture the real implications of systems and how damaging they can be.”.

A significant portion of respondents favoured extrinsic evaluations, with 32% stating this as their preference, highlighting its direct relevance to real-world fairness and discrimination concerns. They emphasised that extrinsic metrics assess how a system behaves in practice and whether it causes harm which many considered of high importance. Context specificity was also noted as crucial—certain biases may be unacceptable in some applications: “For example, in language-vision models, for some contexts there may be associations/stereotypes that are not acceptable (e.g., only generating images of male footballers) and some that are expected/acceptable (e.g., not generating images of white African leaders).”. Extrinsic evaluation was

| | Number Selected |
|---|---|
| Intrinsic | 1 |
| Extrinsic | 6 |
| Both | 12 |

Table 2: Results from the pilot study for preferred bias measure: intrinsic, extrinsic, or both.

seen as essential for ensuring the safety and fairness of deployed models. Only one participant explicitly preferred intrinsic evaluation.

Clearly, there is value in producing both measurements to allow system users to see if both the model itself and the downstream processes are fair, so a bias warning system should be flexible enough to consider intrinsic and extrinsic measures.

### 5.3 Q3: Bias Warnings

The answers to the third question varied widely, with participants highlighting several key approaches. Table 3 shows the range of answers given, which can be summarised as follows.

**Visualisation** was widely preferred, as participants said it could provide an explicit and intuitive way to identify biased patterns in model outputs. Some users felt they would value example-based visualisations, providing clear and insightful information. Others suggested highlighting biased words directly in model outputs as an additional means of raising awareness, using for instance existing tools like the LLM Sandbox.[3]

**Benchmark scores** were frequently mentioned as a valuable way of assessing and comparing bias across different models. These scores were seen as especially helpful for users who may not have the time or expertise to analyse bias in depth. One participant compared this to certification systems like B-Corp, which provide a quick, external validation for businesses adhering to the highest standards of social impact.

**Explainability** was seen as essential by several practitioners advocating for improved methods to clarify how biases emerge in models. Participants emphasised the need for clear explanations of why certain outputs were generated, how input variations affect bias, and where systemic gaps exist. Examples of interpretability tools for language models include ferret (Attanasio et al., 2023) and Inseq (Sarti et al., 2023).

**Caution alerts** were also considered valuable, particularly as a way to warn users when a prompt might trigger biased responses proactively. One participant suggested that, alongside alerts, the system should offer alternative, less biased outputs.

| | Number Selected |
|---|---|
| Caution alert | 2 |
| Visualisation | 8 |
| Data distribution | 2 |
| Benchmark scores | 7 |
| Explanation | 3 |
| Argumentation | 1 |

Table 3: Results from the pilot study for preferred warning type. Note that we allowed participants to choose multiple answers.

**Data distribution** was also found to interest some participants, as seeing statistics on dataset composition, particularly to understand whether the data used to train and/or finetune models was balanced or skewed, was seen as useful.

One participant felt that **argumentation**-based reasoning, where models would provide logical proof for their outputs, would make their decision-making process more transparent, and easier to identify bias within the process.

## 6 Bias Warning Framework

As discussed in Section 5.3, there are some differing opinions on how bias warnings should be reported, but the consensus tends to favour visualisations and benchmark scores. One way to produce benchmark scores and visualisations for each model prediction's bias is to *monitor* the model producing the output. We propose a bias warning framework that leverages ideas from deep neural network monitoring.

**Existing monitoring methods.** Most runtime monitoring literature focuses on misclassification or out-of-distribution detection (Guerin et al., 2023), where a runtime monitor is used to improve the safety of machine learning models by detecting unsafe outputs encountered at inference time. The monitor sits alongside the underlying model. It takes in the same inputs as the model and model outputs to accept or reject an output. Many monitors utilise a scoring method, for example, based on distance (Liu and Qin, 2023), energy score (Liu et al., 2020), or feature importance (Sun and Li, 2022). Recently, Naveed et al. (2024) propose a framework to monitor 'human-centric requirements', where the monitor consists of multiple fair-

[3] https://ai-sandbox.list.lu/

ness metrics, both intrinsic and extrinsic, calculated on the model's output.

**Our framework.** With this in mind, we propose the following monitoring framework for bias warning in language models, depicted in Figure 1. Our bias monitor generates quantitative bias scores by analysing model inputs, outputs, previous model outputs, and previous monitor outputs. This monitor will recompute these scores on an input-by-input basis. In other words, bias is checked for each new input and prediction. This means we can easily extract inputs that produce unfair outcomes for retraining purposes. As discussed in Section 5.1, respondents generally agreed that bias mitigation is best at the data curation stage of language models' development pipeline. By utilising our monitoring framework, practitioners can find the inputs that affect the model's fairness in real time. These inputs can be gathered to retrain the model and thus can help in the data curation step of the development process. Moreover, by allowing previous model outputs to be included, practitioners can also see if bias has changed over time, and can compute bias measures requiring more than one output. Our monitoring framework accounts also for visualisation, which was the preferred bias warning by the practitioners in our pilot study. Indeed, the bias monitor's outputs can be easily incorporated into a visualisation. For example, we can imagine a traffic light system based on thresholds on the various benchmark scores output by the monitor. Ultimately, our bias warning framework is highly customizable as different scoring methods could be added or removed, and these scores can be calculated in a *post-hoc* manner as the monitor will not need to alter the inner workings of these models; they just need access to the outputs. Additionally, as the monitor does not need to be aware of the inner workings of the model, third-party control bodies can configure and use it to increase trust in these systems.

**Example.** To illustrate how our bias monitoring framework might work, we provide an example in Table 4. Suppose we have an AI system like a language model that decides whether to approve or reject bank loans, considering each person's gender, income, and credit rating (low or high). In this example, the monitor calculates the demographic parity and disparate impact of the model outputs for each input and outputs these values to the user. Demographic parity in this case will be calculated as $|P(Accept|Male) - P(Accept|Female)|$, and disparate impact is calculated as $\frac{P(Accept|Female)}{P(Accept|Male)}$. After the first three inputs to the dataset, the monitor will output a demographic parity of 0 and a disparate impact of 1 based on the definitions of these metrics given above, showing no bias present. After the fourth individual, the new demographic parity is 0.5, and the disparate impact is 0.5, indicating that the model may be biased against female applicants. With the addition of the fifth data point, the demographic parity is 0.167, and the disparate impact is 0.75, which is an improvement. With the sixth input, the bias worsens with demographic parity at 0.334 and disparate impact at 0.5. Using this series of monitor outputs, we can determine which inputs may affect the model's bias. In this case, we should consider looking at inputs 4, 5, and 6 more in-depth. This process will be more informative with more complex datasets and more fairness measures.

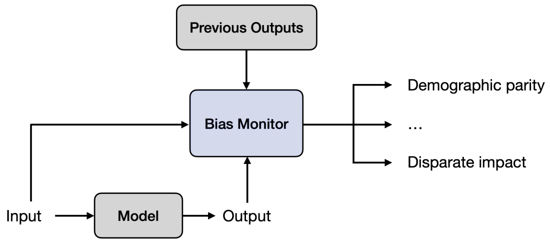

Figure 1: Our proposed bias monitor framework. This monitor takes new inputs to the underlying model, the model outputs, and also previous outputs of both the model and the monitor, and then outputs multiple benchmark scores based on these inputs. This example shows demographic parity and disparate impact, but the monitor can be personalised to account for other bias metrics.

| | Gender | Income | Credit Rating | AI Decision |
|---|---|---|---|---|
| 1 | Male | 50k | Low | Accept |
| 2 | Male | 150k | High | Accept |
| 3 | Female | 200k | High | Accept |
| 4 | Female | 80k | Low | Reject |
| 5 | Male | 80k | Low | Reject |
| 6 | Female | 150k | High | Reject |

Table 4: An example of inputs for our monitoring framework.

## 7 Conclusion

The widespread adoption of language models paired with their socio-demographic biases can perpetuate societal inequalities across many use cases. While substantial efforts in NLP research have been made to measure and mitigate these biases, this research highlights the often-overlooked aspect of how such biases are communicated to practitioners, which instead is a crucial aspect as it can increase user trust and understanding of these models. In this paper, we address this gap by conducting a user study on bias mitigation, measurement and warning in language models, targeting female AI practitioners during a workshop promoting gender-inclusive AI systems. Specifically, we focus on gender bias and further study how practitioners' choices generalise to other socio-demographic biases. Our study reveals that user preferences for bias mitigation and measurement show strong consensus, in contrast to the wide variation in user preferences for bias communication, emphasising the need for tailored approaches of bias warnings. Based on these findings, we develop a user-centred framework for personalised bias reporting integrating runtime monitoring techniques into language models to assess and visualise biases dynamically. Future work can expand on this preliminary framework in several directions to explore its applicability and impact more broadly. For instance, researchers could evaluate the framework using established datasets from AI Ethics research, such as those in the financial domain (Hardt et al., 2016), to better understand how well it supports practitioner workflows. Another promising direction is to conduct a before-and-after user study to assess the framework's potential in fostering user trust in AI systems, following methodologies similar to Di Bonaventura et al. (2024). Overall, this study opens up a range of possibilities for tailoring bias communication strategies and integrating user-centred tools into real-world model deployments.

## Limitations

We are aware of the following limitations. **(1)** The number of responses for the user study is limited; a wider study would be required for more statistically significant results and to draw more robust conclusions. **(2)** The study would benefit from a more diverse set of respondents, both concerning gender and race, but also with different years of experience in machine learning. Moreover, we treated gender as a binary category, i.e., male/female, and disregarded other important categories at their intersection, such as the trans community. Future work should expand this as we anticipate that different groups would have different preferences for bias warnings. **(3)** We focused on assessing individual preferences around gender bias in language models from mitigation and measurement to warning. However, we did not investigate preferences across different applications and domains. This is an interesting direction for future work, as participants in our survey briefly mentioned different preferences across domains and use cases, e.g., the medical domain and machine translation. **(4)** We focused on assessing individual preferences around bias in our pilot study, whose findings we used to develop our personalised bias monitoring framework. As such, respondents were not asked to evaluate our proposed monitoring framework. Future work should explore the proposed bias warning framework in depth by, for instance, collecting user feedback.

## Acknowledgments

This work was supported by the UK Research and Innovation [grant number EP/S023356/1] in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). We acknowledge the Department of Informatics, King's College London, for supporting the initial workshop at ACM womENcourage 2024 through its Athena Swan initiatives. In particular, we are grateful to Alfie Abdul Rahman for her support. We thank Madeleine Waller, Mackenzie Jorgensen, and Martim Brandão for their insightful feedback.

## References

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Joan Giner-Miguelez, Pieter Gijsbers, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Michael Kuchnik, et al. 2025. Croissant: A metadata format for ml-ready datasets. *Advances in Neural Information Processing Systems*, 37:82133–82148.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 256–266, Dubrovnik, Croatia. Association for Computational Linguistics.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Federico Bianchi, Amanda Cercas Curry, and Dirk Hovy. 2023a. Artificial intelligence accidents waiting to happen? *Journal of Artificial Intelligence Research*, 76:193–199.

Federico Bianchi, Tommaso Fornaciari, Dirk Hovy, and Debora Nozza. 2023b. *Gender and Age Bias in Commercial Machine Translation*, pages 159–184. Springer International Publishing, Cham.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901.

Christina Blacklaws. 2018. Algorithms: transparency and accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170351.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Alessandro G Buda, Greta Coraglia, Francesco A Genco, Chiara Manganini, and Giuseppe Primiero. 2024. Bias amplification chains in ml-based systems with an application to credit scoring. *Proceedings of the 3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024*.

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

Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist tools: Social class correlates with performance in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655, Bangkok, Thailand. Association for Computational Linguistics.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. Is explanation all you need? an expert survey on LLM-generated explanations for abusive language detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Bhavana Doppalapudi, Md Dilshadur Rahman, and Paul Rosen. 2024. Seeing is believing: The role of scatterplots in recommender system trust and decision-making. In *International Symposium on Visual Computing*, pages 425–438. Springer.

Laura Douglas. 2017. Ai is not just learning our biases; it is amplifying them. *Medium, December*, 5.

Aimen Gaba, Zhanna Kaufman, Jason Cheung, Marie Shvakel, Kyle Wm Hall, Yuriy Brun, and Cindy Xiong Bearfield. 2023. My model is unfair, do people even care? visual design affects trust and perceived bias in machine learning. *IEEE transactions on visualization and computer graphics*.

Yacine Gaci. 2023. *On Subjectivity, Bias and Fairness in Language Model Learning*. Theses, Université Claude Bernard - Lyon I.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Joris Guerin, Kevin Delmas, Raul Ferreira, and Jérémie Guiochet. 2023. Out-of-Distribution Detection Is Not All You Need. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14829–14837. Number: 12.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022a. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022b. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1078–1088.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Litian Liu and Yao Qin. 2023. Fast Decision Boundary based Out-of-Distribution Detector. *arXiv preprint*. ArXiv:2312.11536 [cs, eess].

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Hira Naveed, John Grundy, Chetan Arora, Hourieh Khalajzadeh, and Omar Haggag. 2024. Towards Runtime Monitoring for Responsible Machine Learning using Model-driven Engineering. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, MODELS '24, pages 195–202, New York, NY, USA. Association for Computing Machinery.

Gabriele Nino and Francesca Alessandra Lisi. 2024. Rethinking bias and fairness in ai through the lens of gender studies. *Proceedings of the 3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024.*

Debora Nozza, Federcio Bianchi, Dirk Hovy, et al. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *arXiv preprint arXiv:2502.20897*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Jon Rueda, Janet Delgado Rodríguez, Iris Parra Jounou, Joaquín Hortal-Carmona, Txetxu Ausín, and David Rodríguez-Arias. 2024. "just" accuracy? procedural fairness demands explainability in ai-based medical resource allocations. *AI & society*, 39(3):1411–1422.

Gabriele Ruggeri and Debora Nozza. 2023. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Association for Computational Linguistics*.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Nasim Sobhani and Sarah Delany. 2024. Towards fairer NLP models: Handling gender bias in classification tasks. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 167–178, Bangkok, Thailand. Association for Computational Linguistics.

Yiyou Sun and Yixuan Li. 2022. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In

*Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 691–708, Cham. Springer Nature Switzerland.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ data science*, 5:1–24.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Marta Ziosi, David Watson, and Luciano Floridi. 2024. A genealogical approach to algorithmic bias. *Minds and Machines*, 34(2):9.