# Sensemaking in User-Driven Algorithm Auditing: A Case Study on Gender Bias in an Image Captioning Model

Behnoosh Mohammadzadeh
Université Paris Saclay, CNRS, LISN
Orsay, France
behnoosh.mohammadzadeh@universite-paris-saclay.fr

Jules Françoise
Université Paris-Saclay, CNRS, LISN
Orsay, France
jules.francoise@lisn.fr

Michèle Gouiffès
Université Paris Saclay, CNRS, LISN
Orsay, France
michele.gouiffes@lisn.fr

Baptiste Caramiaux
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR
Paris, France
baptiste.caramiaux@sorbonne-universite.fr

## Abstract

Non-experts increasingly engage in user-driven algorithm auditing, interacting directly with AI systems to probe, document, and reflect on biased behavior. Yet, auditing remains challenging due to model opacity and limited support for navigating and interpreting outputs. This paper explores the design and evaluation of interfaces grounded in the sensemaking framework to support non-experts in auditing gender bias in image captioning. In a between-subjects study, 60 participants audited an image captioning model using one of three interface conditions: a Baseline interface, a Masking Tool for image manipulation, or a Filtering Tool for organizing captions. Our findings show that interface design shaped what participants noticed, how they interpreted model behavior, and supported their hypotheses. The Image Masking Tool enabled fine-grained testing of visual cues and context, while the Text Filtering Tool revealed broader asymmetries in gendered language. We argue that incorporating sensemaking into auditing practices can advance accountability and transparency in machine learning systems.

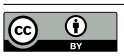
## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → Computer vision.

## Keywords

user-driven algorithm auditing; sensemaking; human-centered AI; algorithmic fairness; image captioning; gender bias

**ACM Reference Format:**
Behnoosh Mohammadzadeh, Jules Françoise, Michèle Gouiffès, and Baptiste Caramiaux. 2026. Sensemaking in User-Driven Algorithm Auditing: A Case Study on Gender Bias in an Image Captioning Model. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3772318.3790784




*CHI '26, Barcelona, Spain*

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790784

## 1 Introduction

People increasingly encounter machine learning (ML) systems in everyday life, from search and recommendation engines to accessibility tools and creative applications. As these systems assume roles with greater social implications, their outputs can shape what users see, do, and believe, often without transparency into how those outputs are generated. In response, users are left to interpret algorithmic behavior through direct interaction, typically without support or explanation. This interpretive work becomes critical when users encounter outputs that appear biased, inaccurate, or otherwise problematic. In recent years, non-experts have brought widespread attention to such issues through everyday engagement with deployed systems. For example, people uncovered racial bias in Twitter's image-cropping algorithm [17] and documented discriminatory search results that exposed systemic profiling [41].

Building on these examples, recent research in HCI and STS has investigated *user-driven algorithm auditing*, an approach where non-experts engage with opaque systems to understand model behavior and surface harms [10, 39]. Unlike expert-led audits, these practices focus on the perspectives of laypeople, revealing issues that can go unnoticed in technical evaluations. At the center of such practices is the interpretive effort users undertake as they probe outputs, compare examples, and generate hypotheses about what the system is doing. This iterative process is referred to as *sensemaking* [10, 34]: an effort to identify patterns, test assumptions, and refine understandings of model behavior.

However, a core challenge lies in how to support this interpretive work. Prior tools have offered new ways for users to engage in auditing, but much of this support has remained oriented toward performance evaluation or guided explanation [21, 33, 46]. In contrast, sensemaking requires more open-ended support: users need ways to notice patterns, compare outputs, and test hypotheses without having their agency constrained or attention biased [8, 33]. Yet, the role of interface design in shaping these processes has not been systematically studied, raising the central research question of this work: how do interface design choices driven by sensemaking processes impact audits by non-expert users?

To investigate this question, we study auditing of profession-related gender bias in an image captioning model by non-expert users. Technical evaluations in this domain have already shown

gender biases, such as models tending to associate certain occupations with specific genders, thereby reinforcing harmful stereotypes [28, 50]. Given the societal impact of how gender and professions are represented, image captioning offers a compelling use case for examining user-driven algorithm auditing through a sensemaking lens. We conducted a between-subjects online study with 60 participants, who were asked to audit the Salesforce BLIP [23] model for occupational gender bias. Participants were assigned to one of three conditions. A Baseline interface allowed open-ended exploration of an image dataset and its generated captions. An Image Masking Tool extended this by letting users draw masks over specific image regions to test how visual content influenced captions. A Text Filtering Tool enabled participants to query the dataset based on words present in the output captions, supporting comparisons across similar instances. We performed a mixed-methods analysis of the audit outcomes. Our findings show that non-expert participants uncovered a range of gender biases in the model's behavior, including the reinforcement of gendered professional stereotypes, a tendency to prioritize gender over profession, a biased reliance on visual cues, and the use of gendered language. Their audits revealed that these biases often emerge through complex interactions between visual and linguistic modalities. The types of hypotheses participants developed were shaped by the auditing tools: the Image Masking Tool enabled fine-grained testing of visual and contextual cues, while the Text Filtering Tool helped reveal broader asymmetries in how men and women were described. This work demonstrates that the design of auditing interfaces influences what patterns of bias users identify, how they interpret model behavior and support their hypotheses.

## 2 Related Work

This section begins by reviewing sensemaking as a key interpretive process through which users engage with algorithmic outputs, then situates this within the broader shift from expert-led to user-driven algorithm auditing, and finally reviews documented patterns of gender bias in image captioning systems.

### 2.1 Background on Sensemaking of Algorithmic Behavior

The Sensemaking framework models how people interpret and organize complex information during exploratory analysis [34]. Developed through cognitive task analyses with intelligence analysts, Pirolli and Card's model identifies four key phases: gathering raw information, representing it in structured ways, manipulating these representations to generate insights, and ultimately producing a knowledge artifact or decision.

In HCI and CSCW, sensemaking was used to guide the design of systems that support exploration and interpretation of complex data. For example, it has informed tools for navigating computational notebooks [5], multilevel exploration of large language model outputs [40], and online research platforms that combine machine assistance with user inquiry [36]. Sensemaking also underpins work in interactive machine teaching, where it helps decompose complex concepts into teachable components [32]. More recently, the framework has been applied to the design of human-AI collaborative auditing systems for large language models. AdaTest++ [37], for example, scaffolds both expert and non-expert investigations by helping users organize test cases, use prompt templates to probe outputs, and iteratively refine mental models of failure patterns.

Building on this earlier work, Cabrera et al.[4] adapted Pirolli and Card's [34] framework to describe how AI practitioners make sense of *model behavior*. Model behavior refers to the patterns, tendencies, or failures exhibited by an AI system, as perceived and understood by practitioners through a process of sensemaking. Their framework characterizes sensemaking as a non-linear, iterative process comprising four stages: *Instances and Outputs*, *Schemas*, *Hypotheses*, and *Assessment* (Figure 1).

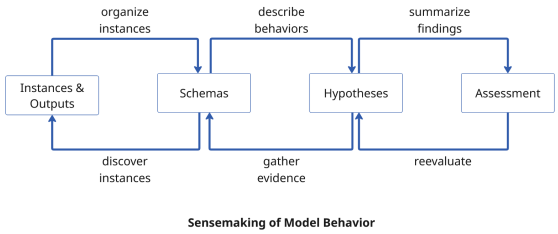



**Figure 1: Adapted from Cabrera et al. [4]. Sensemaking of model behavior is iterative and non-linear: practitioners may enter at any stage, revisit previous steps, and shift between organizing, hypothesizing, and evaluating model behaviors.**

Practitioners begin with *Instances and Outputs*, collecting system outputs from various sources. These are then grouped into semantically meaningful categories in the *Schemas* stage. In the *Hypotheses* stage, they define expectations about model behavior and gather evidence to evaluate them. Finally, the *Assessment* stage involves organizing and reporting findings. Importantly, the framework emphasizes that practitioners may enter at any point, for instance starting with pre-existing hypotheses rather than gathering instances. In addition, practitioners frequently revisit schemas as they refine and update their understanding. Schemas evolve dynamically in response to new outputs and observations, reflecting how practitioners build mental models of model behavior over time.

To evaluate their sensemaking framework, Cabrera et al.[4] designed an analysis tool and conducted a think-aloud study in which data scientists explored, tested, and compared the behavior of two image captioning models.

We are interested in applying the sensemaking framework to the design of interfaces in the context of user-driven AI-based algorithm auditing.

### 2.2 Algorithm Auditing: From Expert-Led to User-Driven Approaches

Algorithm audits involve systematically querying an algorithmic system to observe its outputs and draw conclusions about its functionality, thereby uncovering insights into its opaque inner workings and external impacts [29]. Auditing has been central to diagnosing harmful algorithmic behaviors in real-world contexts, with systematic reviews confirming that discrimination and bias consistently emerge as dominant themes [2, 3, 11, 35, 38, 48]. For example, key studies such as Gender Shades [3] revealed that commercial

facial recognition systems performed far worse on darker-skinned women, while Kay et al. [18] documented the under-representation of women in Google image search. However, most audits remain expert-driven, conducted by researchers and developers with the technical skills required to interrogate systems [10]. This expert focus risks overlooking the lived experiences of end-users, who often encounter harms in everyday contexts [33, 39].

Recent research and online discussions have highlighted how non-expert users can uncover harmful algorithmic behaviors through direct interaction with systems [11, 19, 24, 26, 39]. User-driven audits have the potential to empower non-experts to challenge algorithmic behavior and contribute to broader accountability efforts [7, 22, 27, 39]. Researchers have therefore proposed interactive tools to better support users in auditing algorithms, such as Recast [46] or IndieLabel [21] for content moderation. Closer to our interest, WeAudit [9] was recently introduced as a workflow and platform co-designed with end-users and practitioners to scaffold user engagement in auditing while ensuring findings are interpretable and actionable.

Researchers have also proposed models outlining the cognitive and behavioral processes involved. Shen et al. [39] described 'everyday algorithm auditing' as a process through which users organically detect, document, and share problematic behaviors, often through social media or online communities. Their work highlights a recurring pattern: users notice a biased output, raise awareness, form hypotheses, test responses across examples, and sometimes suggest remediations. Building on this, DeVos et al. [10] offered a three-stage model that formalizes the cognitive processes of auditing into: noticing patterns in outputs (search inspiration), interpreting those outputs to understand possible bias (sensemaking), and considering actions or responses (remediation). Beyond outlining these stages, their model also emphasizes how users' prior knowledge and platform affordances, such as autocomplete, filtering, or search result layouts, influence what users can observe, explore, and compare. Crucially, prior literature observed that sensemaking is not only a point of deep engagement with affordances, but also a site where users' understanding evolves. For example, two main sensemaking strategies in user-led audits have been observed: (1) manipulating single inputs to compare how outputs change, and (2) comparing multiple outputs in parallel to detect broader patterns of bias [17, 41].

While these works have identified the sensemaking strategies non-expert users employ when auditing algorithms, a key challenge remains: how to design tools that effectively support this iterative, interpretive process. In this paper, we examine how different tool designs, grounded in observed sensemaking strategies, shape how non-expert users interpret algorithmic behavior and affect the outcomes of audits.

## 2.3 Gender Bias in Image Captioning Models

This article focuses on the case of auditing gender bias in automatic image description algorithms. Image captioning models often exhibit gender bias, as documented in expert-led evaluations [16, 43]. Here, we propose to present the literature on gender bias in image captions, structuring it around the patterns of bias that these previous studies have identified.

*Stereotype reinforcement* is a recurring pattern wherein models amplify gender associations, such as describing snowboarders as "men" or individuals in kitchens as "women," even when gender evidence is unclear or absent [16, 49]. Mandal et al. [28] further showed that occupations like "beautician" or "housekeeper" are disproportionately associated with women, while roles such as "engineer" are linked to men, correlations that mirror real-world disparities in workforce participation and pay gaps. Nakashima et al. [31] further identified *context-to-gender bias*, where background cues such as kitchens trigger female predictions, and *gender-to-context bias*, where perceived gender leads to hallucinated objects, such as a suit for men. Such behaviors are prevalent in captioning models [31, 43] and become even more pronounced in TTI systems [25].

Other studies reveal deeper structural mechanisms. Abdollahi et al. [1] describe *gender–activity binding*, where changing only the gender of a subject systematically alters predicted roles, even with identical visual inputs. Using the GAB dataset, they show that models perform worse when confronted with role–gender pairings that contradict stereotypical expectations. Xiao et al. [47] extended this with the GenderBias-VL benchmark, demonstrating that occupational inferences can shift solely based on gender perturbations, despite identical visual and contextual cues.

Wolfe and Caliskan [45] identified *markedness bias*, whereby male-presenting individuals are more often labeled with neutral terms such as "person," while female-presenting individuals are explicitly marked with gendered labels like "woman." This asymmetry, visible in models such as CLIP, also shapes downstream generative outputs.

In addition, the persistent use of *binary gender framing* reinforces *misgendering*. Models trained on binary gender classification frequently rely on stereotypical cues from appearance, context, or role, leading to systematic errors [15, 16, 42] and erasing non-binary identities that remain underrepresented in datasets and benchmarks [42].

Finally, the VisoGender benchmark highlights *resolution bias*, where models systematically default to masculine labels when resolving gender in ambiguous contexts. Masculine pronouns are predicted more reliably than feminine ones, and in multi-person images, contextual stereotypes further skew resolution toward men, leading to systematic misclassification [15].

While previous work has identified critical patterns of gender bias in vision-language models, most analyses are based on benchmark experiments and do not result from algorithm audits by users. As a result, little is known about how effectively users can identify such bias patterns. Using gender bias in image captioning model as use case, our study investigates whether interactive auditing systems, grounded in the concept of sensemaking, can support users in detecting these patterns.

# 3 Methods

We conducted an experimental study in which participants audited an image captioning model for gender bias. This section details the study design and experimental conditions, participant recruitment, procedure, and the methods used for data collection and analysis.

## 3.1 Participants

We recruited 60 participants through the Prolific platform[1]. We screened for individuals with no prior experience in algorithm auditing and no formal education in machine learning. Participants ranged in age from 22 to 65 years (M = 36.4), with a balanced gender distribution (31 male, 29 female). The sample was internationally diverse, with participants primarily residing in South Africa, the United Kingdom, Poland, and the United States. Educational backgrounds varied: 22 participants held a Bachelor's degree, 18 held a Master's degree, and the remainder had completed college-level, high school, or other forms of education. Participants reported moderate familiarity with machine learning and image captioning, most frequently rating themselves at 2 or 3 on a 5-point Likert scale, confirming that they were not domain experts.

## 3.2 Image Captioning Model and Dataset

The main task of the study was the auditing of a specific image captioning model. For this study, we selected the Salesforce BLIP model [23], a widely used image captioning system known for its rich, descriptive outputs and relevance in applications such as digital accessibility.

To further frame the task, we focused on gender bias in occupational settings using the Visogender dataset [15], a collection specifically curated to support the study and measure of gender bias in visual-language models. The dataset includes 230 images in the single-person set and 460 images in the two-persons set, covering 23 occupations. It was built with specific inclusion criteria: each image features one or two individuals; images exclude children, stock photo watermarks, and non-safe-for-work content; and all images are publicly available under Creative Commons licenses. Images are either in color or black and white and depict people engaging in occupational or contextual activities. These properties ensured both ethical use and conceptual clarity for participants evaluating the model's outputs.

For this study, we selected a subset of 80 images from the medical sector, where preliminary testing revealed particularly salient examples of gender bias in model-generated captions. Images of medical professionals, such as doctors and nurses, consistently elicited captions that reflected entrenched gender stereotypes across cultural contexts. Figure 2 (left) depicts examples of images from the dataset used in the study.

## 3.3 Auditing interfaces

We designed a web-based interface that scaffolds the key stages of the sensemaking process as described in Cabrera et al. [4] (and depicted in Figure 1), including interpreting instances and outputs, aggregating them into schemas and generating hypotheses, and organizing evidence.

*3.3.1 Auditing interface design.* The interface, displayed in Figure 2, consists of three main components: an image dataset explorer, an image-caption viewer, and a structured documentation panel for reporting biases. The **Dataset Explorer** and **Model's Captions** panels support the *Instances & Outputs* stage by allowing users to scroll through and select images from the dataset, which opens a larger preview in the center, accompanied by its model-generated caption, enabling close reading and comparison of outputs.

On the right, users can create and manage **Bias Cards**, which serve as structured units for supporting both the *Hypotheses* and the *Schemas* stages of the framework. Each card includes a text field where users describe a suspected bias and articulate their hypothesis. Below this, a drag-and-drop area allows users to compile supporting evidence by selecting relevant image-caption pairs. This visual organization of related examples supports the creation of *Schemas*, helping users iteratively group semantically similar behaviors and refine their understanding. Users can revise cards as their hypotheses evolve, and multiple cards can be created to represent different observations.

At the end of each session, all cards are compiled into a final audit report that reflects a user-generated *Assessment* of model behavior.

*3.3.2 Auditing tools.* Because the audit task involves examining the relationship between images and their textual descriptions, we developed two additional tools specialized for each modality: an **Image Masking** tool and a **Text Filtering** tool. The tools were designed to support different sensemaking strategies identified by DeVos et al. [10]: manipulating single inputs to compare changes in the output, and grouping outputs to detect broader patterns of bias.

**The Image Masking Tool** allows users to selectively occlude parts of an image with a brush interface and observe how the model-generated caption changes in response (see Figure 3). This design enables users to go back to the *Instances & Outputs* stage [4] and create specific synthetic instances to test their hypotheses. The design of this tool was inspired by strategies observed in user-driven audits. For example, Twitter users tested hypotheses about racial bias in the platform's image-cropping algorithm by altering features such as background color or clothing [17], and Booking.com users adjusted input ratings to uncover patterns of score inflation [39].

**The Text Filtering Tool** enables users to input keywords and dynamically filter the dataset based on terms found in model-generated captions (see Figure 4). Its design was inspired by the *Schemas* stage of the sensemaking process [4], which emphasizes grouping related instances to support broader generalizations. Filtering provides a way for users to move from isolated examples toward identifying recurring patterns across the dataset. Similar strategies underpin prior audits of representational bias: users analyzing Google Images, for example, revealed gender imbalances in query results for "doctor" versus "nurse" [10], while Sweeney's audit of search ads exposed racial discrimination by systematically comparing names at scale [41]. In our interface, users can enter one or more keywords (combined using AND logic) to narrow the dataset view to images whose captions contain all specified terms.

*3.3.3 Implementation.* The auditing interface was developed using Marcelle,[2] a JavaScript toolkit for building interactive machine

[1] https://www.prolific.com/

[2] https://marcelle.dev/

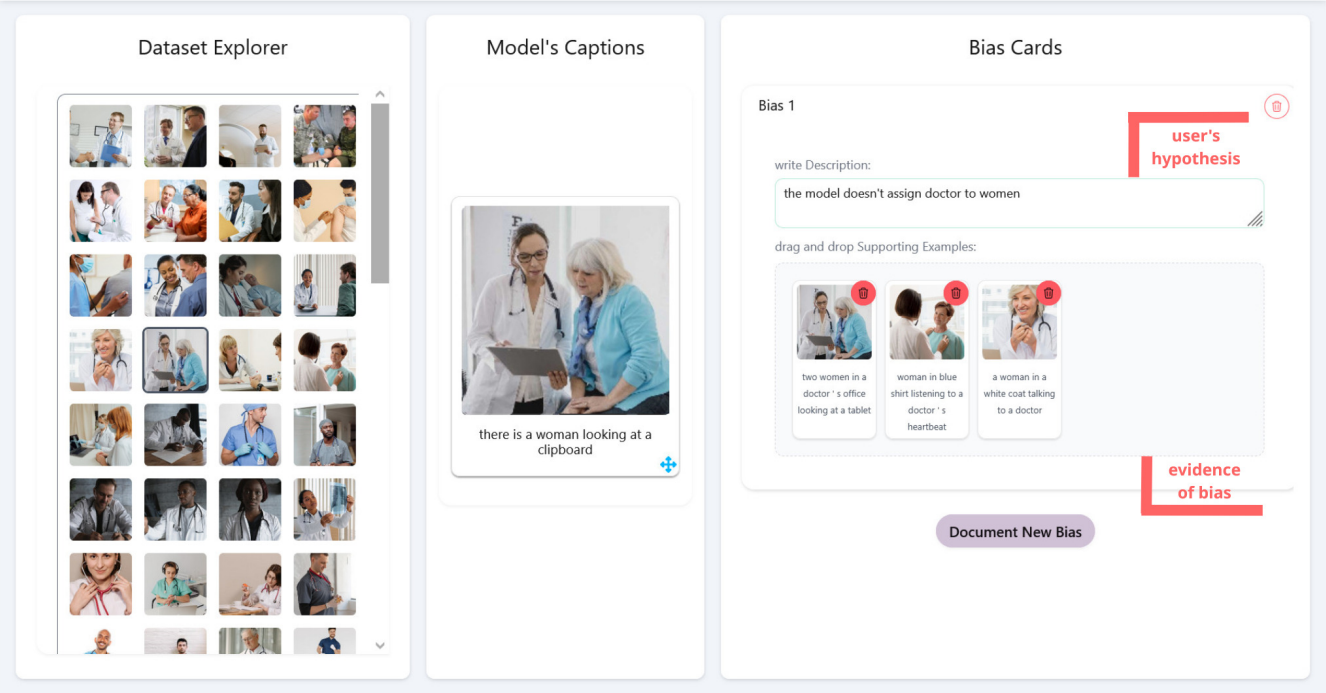



**Figure 2: The Auditing Interface includes: a Dataset Explorer (left), Caption Viewer (center), and Bias Cards panel (right) where users document a description of a the bias (hypothesis) and attach supporting examples by selecting relevant image-caption pairs (evidence of bias)**

learning applications [13]. Marcelle facilitates the integration of ML data and models in web applications and the collection of user data. Its Node.js backend provides user authentication and persistent data storage using MongoDB, and the front-end was implemented using SvelteKit [3]. The auditing interface extends Marcelle with several custom components designed specifically to scaffold user-driven algorithm auditing. In particular, we implemented new modules for the Bias Cards, the Image Masking Tool, and the Text Filtering Tool. The source code for the auditing interface is available as open source for reproducibility [4].

## 3.4 Study Design

The study follows a between-participants design with three conditions:

**baseline**: auditing with the **Baseline** interface (original interface design, Figure 2)

**masking**: auditing with the baseline interface augmented by the **Image Masking Tool** (Figure 3)

**filtering**: auditing with the baseline interface augmented by the **Text Filtering Tool** (Figure 4).

Participants were evenly assigned across the three experimental conditions (n = 20 per condition). All participants were compensated at an hourly rate of $11. This rate aligns with minimum wage in France, is considered fair by Prolific, and helps maintain participant engagement and data quality. Participants were informed that their audits would be reviewed for alignment with task instructions and quality before approval. The study was reviewed and approved by the institution's Research Ethics Committee.

## 3.5 Procedure

Each session lasted approximately 60 minutes and followed a structured format. Participants began with a brief introduction outlining the study's goals, followed by a pre-study questionnaire that collected demographic information and administered the Ambivalent Sexism Inventory (ASI) [14]. The ASI is a validated instrument commonly used in studies of gender bias and was included to assess participants' underlying attitudes toward gender. Prior research suggests that such attitudes may influence how individuals perceive and report algorithmic biases [10]. All questionnaires are included in the supplementary materials.

Participants were then introduced to the auditing interface through a video tutorial and a step-by-step hands-on guide. The tutorial covered the interface's main components and demonstrated how to use the Bias Cards to document findings. Depending on the experimental condition to which they were randomly assigned (described in Section 3.4), the tutorial also introduced either for the baseline, masking, or filtering conditions.

Following the tutorial, participants completed a 30-minute auditing task. They were instructed to explore the dataset, examine the model-generated captions, identify potential examples of gender

[3] https://svelte.dev/
[4] https://github.com/l3ehfar/VLM-Auditing-Interface

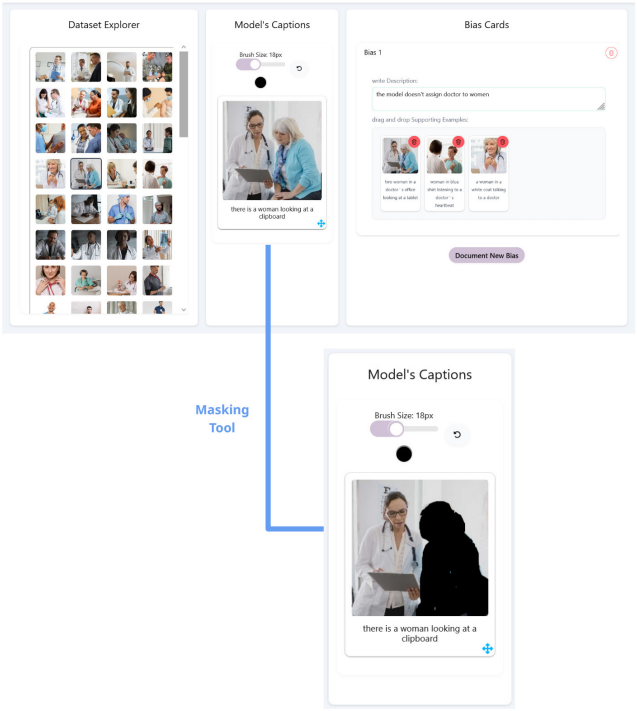



**Figure 3: The Image Masking Tool allows users to test the impact of visual features by selectively masking regions of the image and observing changes in model output.**

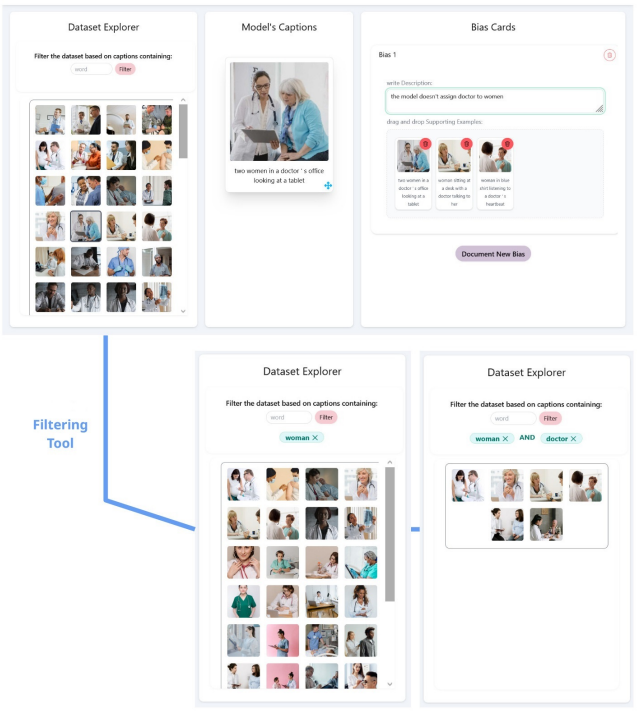



**Figure 4: The Text Filtering Tool allows users to narrow the dataset based on keywords found in model-generated captions. The bottom row illustrates two examples of filtering using multiple keywords with the AND operator, for instance, filtering for images whose captions include both "woman" and "doctor."**

bias, and document as many biases as they found within the allotted time using the Bias Cards, with each card representing a distinct observation or hypothesis.

## 3.6 Data Collection

To investigate how interface design shaped participants' sensemaking during user-driven algorithm auditing, we collected both quantitative and qualitative data throughout the study.

*3.6.1 Demographics and Background.* We collected demographic information (age, gender, country of residence) directly through the Prolific platform. A pre-study questionnaire additionally captured participants' education level, professional field, self-assessed familiarity with artificial intelligence (AI) and image-captioning models, and ASI responses.

*3.6.2 Bias Cards as Sensemaking Artifacts.* During the study, participants created audit reports by organizing biased examples into Bias Cards. These cards included a textual description of the bias they identified (hypothesis), as well as a set of supporting image-caption pairs selected as evidence. After submitting each card, participants completed a brief self-assessment indicating how confident they were in the validity of their hypothesis. We define *confidence* as participants' self-reported certainty in their interpretation of the model's behavior on a 5-point Likert scale. Participants could also add optional comments to elaborate on their reasoning or reflect on the card creation process.

## 3.7 Data Analysis

Our analysis examines how interface design shaped participants' auditing outcomes, combining quantitative comparisons with thematic analysis. We analyze how each tool influenced the number of hypotheses, the amount of supporting evidence, and participants' confidence, as well as the types of gender bias they identified and how they framed visual, linguistic, and cross-modal patterns.

We analyzed a total of 274 valid bias cards, after discarding 51 cards (baseline: 18, masking: 20, filtering: 13) that were either vague, unrelated to gender bias, or did not contain a clear hypothesis. Across these, participants provided a total of 1,090 evidence items (image–caption pairs) to support their hypotheses.

*3.7.1 Quantitative Analysis.* To evaluate the sensemaking outcomes, we used the following indicators:

- *Number of bias cards per participant.* This served as a proxy for the breadth of their bias identification efforts.
- *Number of evidence items per hypothesis.* For each bias card, we counted the number of image-caption pairs that participants grouped as supporting evidence. This metric reflects how thoroughly participants supported their hypotheses.
- *Analysis of confidence and self-assessment questions.* We analyzed participants' self-reported confidence ratings using descriptive statistics and condition-level comparisons.

- *Correlation between confidence and evidence.* This metric provides insight into how grounded participants' certainty was in empirical observations.

Given that most quantitative measures violated normality assumptions (Shapiro–Wilk tests), we used non-parametric analyses throughout. Differences across conditions were evaluated using Kruskal–Wallis tests, for which we report effect sizes using $\eta^2_H$ together with 95% bootstrap confidence intervals. We then used pairwise Dunn tests with Holm correction as post-hoc analysis. Effect sizes for pairwise comparisons are reported using Cliff's $\delta$ with 95% bootstrap confidence intervals. Associations between confidence and evidence were examined using Pearson correlations computed separately per condition; we report correlation coefficients with 95% confidence intervals.

*3.7.2 Qualitative Analysis of Bias Cards.* We conducted a reflexive thematic analysis of participants' bias cards. The first author coded the entire dataset, while the second and fourth authors each independently coded half of the cards. We initially assigned descriptive codes to capture the specific gender biases participants described in their hypotheses (written descriptions: e.g., occupational stereotyping, gender misidentification, omission), using evidence items to disambiguate or assist the interpretation. We aimed to understand how participants made sense of the model's behavior: what biases they identified, how they described them, and how they tested hypotheses regarding the model's behavior. From these initial codes, we collaboratively developed a codebook that defined each code and included example excerpts. The full dataset was then double-coded by two researchers using the finalized codebook, and all four authors participated in resolving disagreements through iterative discussion until consensus was reached.

Eventually, we grouped related codes into broader inductive themes that reflected recurring patterns in what participants identified and hypothesized about gender bias. These themes serve as the basis for understanding how sensemaking outcomes differed across interface conditions. When reporting results, we attribute quotes to participants using anonymized identifiers in the format P<condition>-<ID> (e.g., *PB-13* for participant 13 in baseline condition), to preserve anonymity while indicating their experimental condition.

*3.7.3 Comparative Analysis Across Conditions.* To understand how different interfaces shaped sensemaking outcomes, we conducted both quantitative and qualitative comparisons across the three experimental conditions.

- **Quantitative comparisons** assessed how many times each theme and sub-theme appeared in bias cards related to each condition. We used *chi-square tests* of independence to assess whether the distribution of themes and sub-themes varied significantly by interface condition. Chi-square analyses of thematic distributions include Cramer's $V$ as the effect-size estimate. This approach allowed us to determine whether certain forms of gender bias were more likely to be identified under specific interface conditions.
- **Qualitative comparative analysis** focused on understanding how the design of the interfaces shaped the audit outcomes. For each experimental condition, we examined the bias cards that participants generated across themes and sub-themes. We then aggregated them to identify overarching sensemaking strategies, considering three key dimensions: (1) the nature of the biases described, (2) whether participants explicitly mentioned using the interface when describing the bias, and (3) the evidence supporting their hypotheses (including, in the masking condition, images with manually drawn masks).

## 4 Results

Before presenting our results, we assess whether participants' attitudes towards gender varied across conditions by analyzing ASI scores collected prior to the audit task. One-way ANOVA tests did not show any significant differences across groups for *hostile sexism* ($F(2, 57) = 0.38$, $p = .69$, $\eta^2_p = .01$, 90% CI [<.01, .05]), *benevolent sexism* ($F(2, 57) = 0.38$, $p = .69$, $\eta^2_p = .01$, 90% CI [<.01, .05]), and total ASI score ($F(2, 57) = 0.33$, $p = .72$, $\eta^2_p = .01$, 90% CI [<.01, .05]). Thus, participants' pre-audit ASI scores were comparable across conditions and did not influence the observed differences in auditing outcomes. In this section, we start by reporting a thematic analysis of participants' bias cards. Then, we analyze how different auditing tools influenced participants' sensemaking outcomes, focusing on the distribution of themes across conditions, the quantity of evidence and self-reported confidence.

### 4.1 Thematic Analysis of Participants' Bias Cards

Our thematic analysis of participants' bias cards yielded four interrelated themes that characterize how patterns of gender bias manifested in the model's captions: *reinforcement of gendered professional stereotypes, prioritization of gender over profession, biased reliance on visual cues,* and *language that reflects gendered hierarchies.* These themes do not represent mutually exclusive categories, but rather highlight distinct mechanisms that sometimes co-occurred or built upon one another in participants' audits.

Each theme is presented below with sub-themes and representative quotes from participants' observations. The themes and sub-themes are summarized in Table 1.

*4.1.1* ***Theme 1 – The model reinforces gendered professional stereotypes.*** Participants across all conditions identified patterns in how the model assigned professional roles, often in ways that reflected stereotypical gender norms and undermined women's professional identities.

***T1.1: Men are doctors, Women are nurses:*** Across conditions, participants observed that *"women were rarely described as doctors despite clear indicators of their roles" (PB-17),* such as white coats or stethoscopes, while men were labeled as doctors *"although there are no visual references of him being a doctor" (PM-5).* Some linked this pattern to broader gendered occupational norms: *"It reflects and reinforces the stereotype that nursing is a female-only profession, which may contribute to ongoing gender inequality"* (*PB-10*).

***T1.2: Women's Professional Identity Is Undermined or Replaced.*** Participants reported that women's professional identities were frequently omitted, labeled vaguely, or replaced with passive

**Table 1: Summary of themes resulting from the thematic analysis of participants' bias cards.**

| Theme | Subtheme | Example |
|---|---|---|
| **Theme 1:** Reinforcing gendered professional stereotypes | T1.1 Role labeling (e.g., doctor vs. nurse) | Men labeled as doctors, women as nurses despite identical context or attire |
| | T1.2 Role replacement with domestic context | Woman in uniform labeled as being in a kitchen; visual context downplayed or misread |
| **Theme 2:** Gender foregrounded over profession | T2.1 Gendered labeling | Women labeled as "female doctor" or "woman," while men are simply labeled "doctor" |
| | T2.2 Gender misclassification from appearance | Hair or legs lead to misgendering; when gender is occluded, captions default to male |
| | T2.3 Authority attributed to men in group scenes | In mixed-gender scenes, male is labeled doctor even when woman performs caregiving task |
| **Theme 3:** Visual cues reinforce stereotypes | T3.1 Same clothing, different roles | Male and female in lab coats labeled as "doctor" and "woman" respectively |
| | T3.2 Removing cues changes role attribution | Removing uniform from woman leads to "patient"; man retains "doctor" label without cues |
| **Theme 4:** Language reinforces gendered assumptions | T4.1 Role marked by gender | Use of "female doctor" and occasional "male doctor"; men as default for "doctor" |
| | T4.2 / T4.3 Gendered descriptors diminish women's authority | "Smiling woman": gendered and affective terms more common for women |

or domestic roles. As *PF-2* noted: *"There are women who are clearly health care practitioners or doctors identified as patients or generically"*. In some cases, captions assigned women unrelated domestic contexts; for example, *PM-6* observed: *"Three very similar images, but when there is a woman it provides the wrong role and also the wrong context (kitchen)"*.

#### 4.1.2 ***Theme 2 – The model's emphasis on gender leads to and reinforces stereotypical role assignments***.

Participants across all conditions reported that the model prioritized gender over professional roles, particularly for women, often relying on stereotypical visual cues. They noted that in mixed-gender scenes, professional authority was disproportionately attributed to men.

***T2.1: The Model Describes Women by Gender, Not by Professional Role.*** Across all conditions, participants noted that the model often prioritized gender over professional identity, particularly for women, even when professional cues were visually present: *"Seeing a female doctor wearing her uniform and just calling her a female and not acknowledging her work title is minimizing her position"* (*PB-20*). Others described captions that *"reduce women to just their gender,"* while men were more often labeled by profession (e.g., "doctor"). Several interpreted this asymmetry as more than a technical oversight, suggesting deeper assumptions that *"reinforces real-world gender hierarchies"* (*PM-16*).

***T2.2: Gendering Is Error-Prone, Binary, and Relies on Visual Stereotypes.*** Participants found that the model frequently inferred gender inaccurately, relying on stereotypical cues: *PM-20* wrote: *"Men and women are identified by hair, face and breasts"*; and *PM-8* noted that: *"Because of putting out their leg, the model think is a woman"*. Some participants also critiqued the binary framings: *"Captions are not gender neutral"* and *"There is a possibility of the people being part of LGBTQ"* (*PM-9*).

***T2.3: In Mixed-Gender Contexts, Authority Skews Toward Men.*** Across all conditions, participants observed that when men and women appeared in the same image, the model often failed to attribute professional authority evenly. Groups of men were frequently described as doctors regardless of context: *PB-2* wrote: *"[The model] believes every man in the picture is a doctor, although it is very clear that one of them isn't"*, while *PB-18* commented: *"These captions reinforce male dominance in positions of authority or expertise, particularly in the medical field"*. Conversely, women were misidentified or portrayed in passive roles, when alongside men: *"The model overlooks the professional role of the woman and assumes she is a patient too" (PM-5)*. In some cases, the model reversed the caregiving roles entirely: *"A female doctor assisting a male patient with a vaccine was captioned in reverse"* (*PM-8*).

#### 4.1.3 ***Theme 3 – The model's use of visual cues reinforces stereotypes***.

Participants across all conditions noted that the model relied heavily on visual cues, such as clothing and perceived gender markers, when assigning professional roles. These cues were not interpreted consistently: the same visual signals often led to different role labels depending on perceived gender.

***T3.1: Visual Cues Are Interpreted Differently Based on Gender.*** Across all conditions, participants found that identical visual cues, such as lab coats, scrubs, or stethoscopes, led to different role attributions depending on perceived gender. Women's professional attire was often downplayed or mischaracterized. For instance, *"It is implying that a woman's uniform is belittled to a white shirt"* (*PM-13*); *"The model described a lab coat worn by a woman as a 'white gown'… it sounds like it describes a lady who went to prom rather than a*

*medical professional"* (*PF-14*). In contrast, men's non-professional clothing was sometimes elevated: *"[The model is] assuming that a man is wearing a lab coat when it's a shirt."* Some participants conducted direct comparisons to confirm the discrepancy. *PM-17* wrote: *"Two images of people dressed in the same uniform with stethoscopes, male is referred to as doctor, female is noted as a woman".*

***T3.2: Masking Gender and Context Changes Role Descriptions.*** Participants in the masking condition tested how the model inferred professional roles when gender or contextual cues were obscured. Several reported hiding features such as faces or hair shifted captions from passive or domestic to professional roles. One masked a woman's face and wrote: *"When gender cannot be identified, [the model] changes caption to 'doctor'"* (*PM-15*). Participants also tested how removing contextual cues like uniforms or instruments influenced role attributions: *"When you remove the uniform from the female doctor it says that she is a patient but doesn't do that when the male has no uniform on"* (*PM-3*). In a mixed-gender scene, masking the woman's face led the model to describe the male patient as the doctor: *"though he's clearly the one being treated."*

#### 4.1.4 *Theme 4 – The model's language reinforces gendered assumptions.*

Across all conditions, participants hypothesized that the model's language reinforced gendered assumptions through specific word choices, tone and framing. Rather than merely flagging problematic words, many reported how language framed certain roles as default and subtly reinforced authority hierarchies.

***T4.1: Gendered Adjectives Reinforce Professional Stereotypes.*** Participants across all conditions frequently noted that gendered adjectives (e.g., "female doctor," "male nurse") marked deviations from stereotypical norms. *"[The model] only specifies gender when the doctor is a woman... This implies that being a doctor is inherently male unless stated otherwise"* (*PB-10*); *"By emphasizing the female doctor's gender, the caption reinforces the notion that women in medicine are exceptional, whereas men are the default"* (*PF-20*). Conversely, the term "male nurse" reinforced stereotypes of nursing as a feminine role.

Some participants also noted that gendered adjectives elevated male authority. *"The model emphasizes 'male doctor' when 'doctor' would have sufficed,"* suggesting this framing may *"diminish the perception of women as equally capable in medical roles"* (*PB-7*).

***T4.2: The language Used to Describe Women Lacks Professionalism.*** Participants reported that the model often described women using vague or unprofessional terms rather than occupational titles. *PF-8, PF-10, PF-16, and PB-13* listed labels such as "*white shirt, headphones, toy*," calling them *"belittling when the woman is clearly doing a professional job." PB-5* linked this pattern to broader societal expectations: *"This bias suggests women should be at home doing chores."*

In mixed-gender scenes, participants also noted that women were often described in secondary or passive roles. *PF-1* highlighted sentence structures that minimized women's agency: *"The woman-belittling sentence structure, men being treated or helped 'by a woman', prioritizes a man as someone more important."* Others emphasized how captions misrepresented women's professional presence: *"The caption seems to indicate that the man is far more involved or important in the image than is actually true. The woman is constantly not given an accurate description of her position." (PF-18); "it stated that the woman was looking out of the window but she was clearly examining something to do with her job" (PM-11); and "mistakes the service that a professional provides based on gender" (PF-6).*

***T4.3: Women Are Described by Appearance, Men by Role or Action.*** Participants observed that the model frequently described women by their appearance, expression, or clothing, while men were more often described by professional role or activity. *"The female doctors are labeled as 'smiling female doctor' while male doctors are just stated as male doctors... makes it seem that female doctors have to always have a pleasant expression"* (*PB-20*).

Several participants hypothesized that in professional settings, women were labeled by attire rather than role: *"A woman in white coat talking to a doctor reflects gender bias by implying that the woman in the white coat is not a doctor herself"* (*PB-7*). Comparing two nearly identical images, *PF-18* observed: *"The one says 'a male doctor,' and the other says 'the woman with a robe'"*. Others pointed out a broader pattern: *"More attention is being paid to what a woman is wearing than what a man is wearing"* (*PF-13*).

## 4.2 Effects of the Interface on Auditing Outcomes

This section analyzes how interface design shaped participants' auditing outcomes, combining quantitative comparisons with thematic analysis. We examine how different tools influenced the number and patterns of gender bias participants identified, the amount of supporting evidence they provided, how they framed visual, linguistic, and cross-modal patterns, along with their reported confidence.

#### 4.2.1 *Effect of Interface on Identified Patterns of Bias.*

The distribution of bias cards across the four themes of our thematic analysis according to the interface is depicted in Figure 5. A chi-square test of independence revealed a significant association between interface and theme distribution ($\chi^2(6) = 19.30$, $p < 0.01$, Cramer's $V = .19$), with each tool scaffolding attention toward specific pattern of bias. The masking condition supported visual testing of model behavior, prompting more hypotheses related to Theme 3 (*Visual cues reinforce stereotypes*). The filtering condition enabled linguistic comparison across cases, leading to more hypotheses under Theme 4 (*Language reinforces gendered assumptions*). The baseline condition also led to the highest number of cards in Theme 4 (*Language reinforces gendered assumptions*), as well as in Theme 1 (*Gendered professional stereotypes*).

To examine these differences in greater depth, Table 2 presents the distribution of bias cards across the sub-themes. A chi-square test of independence also revealed a significant association between interface and sub-theme distribution ($\chi^2(18) = 46.66$, $p < .001$, Cramer's $V = .30$).

To gain further insights into participants' use of the auditing tools, and their effect on the patterns of biased identified, we report an analysis of each interface condition. For each condition, we report evidence of the use of the various tools extracted from the thematic analysis, and discuss their impact on the patterns of bias.

We now examine each interface condition in detail:

**Table 2: Number of bias cards per sub-theme and condition.**

| Sub-theme | baseline | masking | filtering |
|---|---|---|---|
| **Theme 1 – Gendered professional stereotypes** | | | |
| Men are doctors, women are nurses | 16 | 21 | 14 |
| Downgrading or omitting women's professional identity | 19 | 7 | 19 |
| **Theme 2 – Gender over profession** | | | |
| Gender prioritized over profession | 16 | 13 | 18 |
| Binary and error-prone gendering | 2 | 11 | 7 |
| Authority skewed toward men in mixed-gender scenes | 4 | 15 | 9 |
| **Theme 3 – Visual cues and stereotypes** | | | |
| Visual cues interpreted differently by gender | 3 | 3 | 3 |
| Masking Gender and Context Changes Role Descriptions | 0 | 7 | 0 |
| **Theme 4 – Gendered language assumptions** | | | |
| Gendered adjectives reinforce stereotypes | 11 | 5 | 8 |
| Language used to describe women lacks professionalism | 3 | 0 | 6 |
| Describing women by appearance, men by role or action | 11 | 5 | 12 |

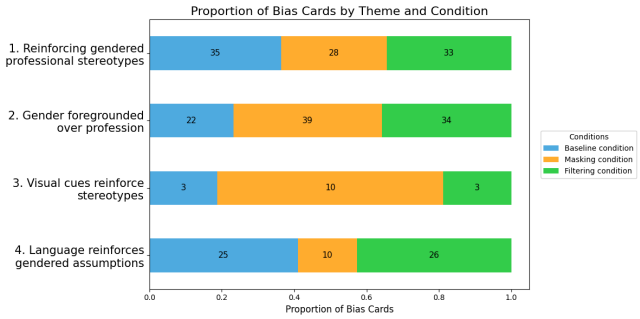



**Figure 5: Proportion of bias cards within each theme by condition. Counts displayed on each bar represent the number of bias cards produced in that condition. This visualization highlights which interface features led participants to focus on which patterns of bias.**

*baseline: Intuitive Observations, Broad Coverage.* In the baseline condition, participants explored captions freely, identifying both isolated anomalies (e.g., misgendering) and broader patterns. While they engaged in the full sensemaking arc, their observations were often grounded in ad hoc comparisons rather than systematic testing. This approach led to the highest number of cards under certain sub-themes in Theme 4, such as *Gendered adjectives reinforce stereotypes* (11 cards), as well as in Theme 1, where *Downgrading or omitting women's professional identity* appeared 19 times. These results indicate that even without targeted tools, participants surfaced diverse patterns of bias through open-ended exploration.

*masking: Focus on Visual Cues and Role Attribution.* The Masking Tool enabled participants to manipulate input images by selectively occluding features such as faces, uniforms, or instruments. This led to fine-grained tests of how visual salience and contextual cues shaped gendered role assignments.

As shown in Figure 6, masking a woman's face in a mixed-gender medical scene often led the model to reassign the role of "doctor" to the man, even when she was holding the stethoscope. Participants' findings in this condition emphasized the interdependence of gender inference and role labeling. These iterative tests resulted in a larger number of hypotheses under Theme 3, particularly the sub-theme *Masking gender and context changes role descriptions* (7 cards), which did not appear at all in other conditions as it built upon the effects of occlusion (Table 2).

Beyond Theme 3, this condition also surfaced high counts in Theme 1 (*Men are doctors, women are nurses*: 21 cards), where several participants used the Masking Tool to occlude facial features and found that when gender cues were hidden, the model defaulted to labeling individuals as male. *"The captions identified the nurse as a male after covering her face"* (*PM-7*). Similarly, *PM-12* observed that a doctor's perceived gender flipped based on face visibility: *"When their face is covered, they are captioned as a man, and when it's not, they are captioned as a woman."*

Finally this condition also surfaced Theme 2 sub-themes, particularly *authority skewed toward men in mixed-gender scenes* (15 cards). In Theme 2, they tested how role attributions shifted when gender cues were selectively occluded: *"The model captions the man as a doctor when the woman's face is covered, even though it is clear that the one receiving treatment is the male and the female is the doctor" (PM-12).*

*filtering: Exposing Linguistic Asymmetries.* The Filtering Tool guided participants to retrieve and compare subsets of image-caption pairs using gender-related keywords. This systematic filtering made it easier to spot linguistic asymmetries such as the disproportionate use of gendered adjectives such as "*female doctor*" versus "*doctor*" alone. Participants also noted unexpected uses of "*male doctor*," which, instead of neutralizing bias, appeared to amplify male authority. As shown in Figure 5 and Table 2, the filtering condition yielded the highest number of hypotheses under Theme 4, especially for *Describing women by appearance, men by role or action* (12

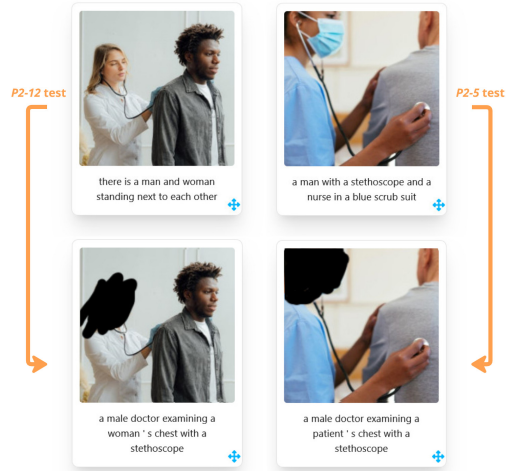



**Figure 6: Example from the masking condition showing how masking the woman's face affects model interpretation. In both cases, the woman is the one using the stethoscope, yet the model attributes the role of "doctor" to the man when the woman is masked.**


**Figure 7: Example from the filtering condition, where a participant filtered the dataset for captions containing both "woman" and "doctor."**

cards) and *Language used to describe women lacks professionalism* (6 cards). Therefore, the tool's focus on filtering from the output caption allowed participants to frame language as not just descriptive, but also as a mean for implicit assumptions about gender and competence.

Moreover, the Filtering Tool enabled participants to quickly verify hypotheses related to occupational labeling, by visually probing what subsets of image produce certain keywords. For example, they reported observing recurrences in Theme 1: *"The word nurse frequently appeared on a caption that identified a woman as a nurse, creating a bias that women should be nurses"* (*PF-16*). They were able to observe that *"The term 'doctor' is only attributed to men"* (*PF-12*), based on the images that remained after filtering for captions containing the word "doctor." Because it aggregates outputs together, the filtering tool helped confirm patterns across the entire dataset: *"The words kitchen and haircut only appear when a female is in the image… no man had such captions"* (*PF-16*). In settings involving several people, participants used this tool to explore gender–role pairings at scale. As shown in Figure 7, retrieving multiple instances with recurring terms (e.g., "doctor" + "woman") enabled them to identify: *"Whenever the two words appear, it is assumed that the woman is not the doctor"* (*PF-16*).

In summary, the Baseline interface enabled participants to observe a broad range of biases through open-ended exploration. The two specialized tools further enabled participants to test more granular hypotheses. The tools effectively supported different sensemaking strategies: instance-level manipulations to create counterfactuals using the Masking Tool, and instance aggregations to validate the significance of linguistic patterns using the Filtering Tool. These led the participants to uncover more diverse patterns of gender bias, focusing either on the effect of visual cues, language patterns, and their interactions.

*4.2.2 Analysis of the amount of bias cards and evidence.* On average, participants submitted 5.55 cards in the baseline condition (SD = 2.76), 5.35 in the masking condition (SD = 2.37), and 5.25 in the filtering condition (SD = 3.08). A Kruskal–Wallis test did not show any significant effect of the condition on the number of bias cards submitted per participant ($H = 0.38$, $p = .83$, $\eta^2_H = .00$).

The amount of evidence per hypothesis according to the condition is reported in Figure 8. A Kruskal–Wallis test revealed a significant effect of the condition ($H = 9.84$, $p < 0.01$, $\eta^2_H = .02$, 95% CI [.001, .073]). Post-hoc Dunn's tests (Holm-corrected) indicated that participants in the filtering collected significantly more evidence per hypothesis than those in the masking condition ($p < .01$, Cliff's $\delta = -.25$, 95% CI [−.40, −.10]).

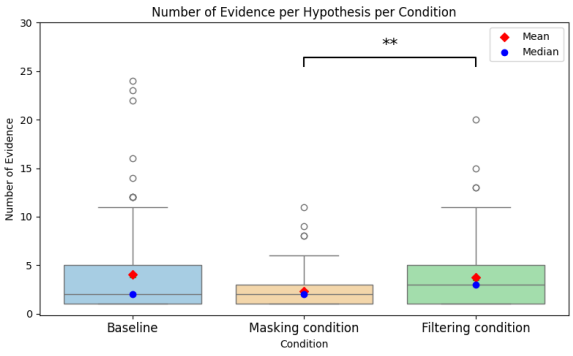



**Figure 8: Distribution of number of evidence items per hypothesis across conditions. Participants in the filtering condition added significantly more evidence per hypothesis than those in masking condition.**

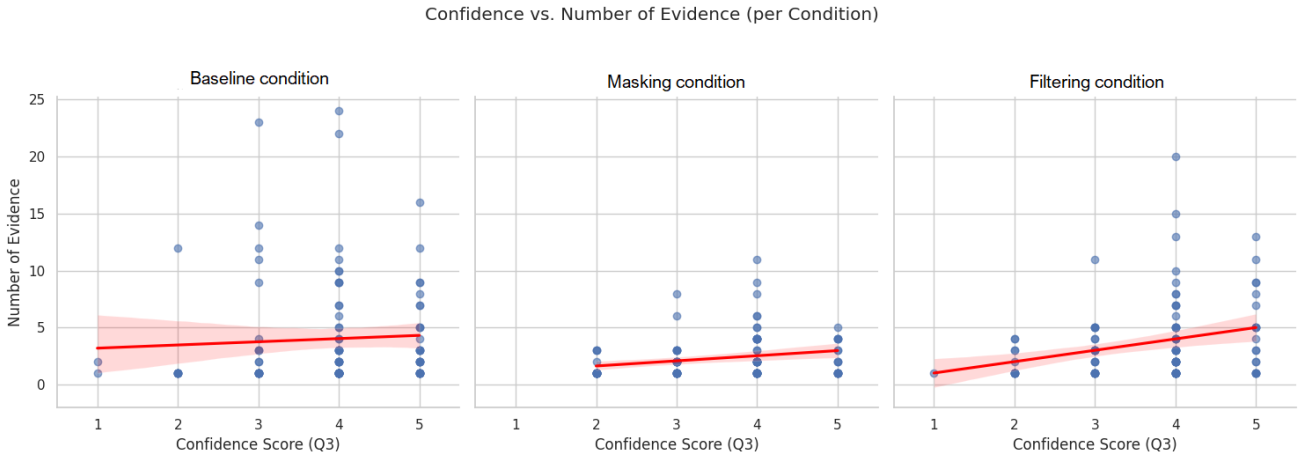



**Figure 9: Correlation between confidence scores and number of evidence items per condition. Significant positive correlations were observed in tool-enabled conditions.**

*4.2.3 Analysis of Confidence ratings.* We analyzed participants' self-reported confidence in their hypotheses across interface conditions. A Kruskal–Wallis test revealed a significant effect of interface ($H = 8.93$, $p < .05$, $\eta^2_H = .02$, 95% CI [.001, .073]). Post-hoc Dunn's tests (Holm-corrected) showed that participants in baseline condition reported significantly higher confidence than those in the masking condition ($p < .01$, $\delta = -.25$, 95% CI [−.40, −.10]). This difference may reflect how the design of the Masking Tool appeared to prompt a more time-intensive, incremental, and exploratory process, as participants revisited initial hypotheses, tested individual modifications to images, and adjusted their conclusions in response to uncertain or conflicting feedback. This suggests that lower confidence may have stemmed from ongoing inquiry.

To examine whether participants' self-reported confidence was related to how much evidence they collected, we conducted Pearson correlation analyses for each condition. In the filtering condition, there was a significant positive correlation between confidence and the number of evidence items per hypothesis ($r = 0.27$, 95% CI [.09, .44], $p = .0049$, $N = 105$). A similar, slightly weaker effect was observed in the masking condition ($r = 0.24$, 95% CI [.05, .41], $p = .014$, $N = 107$). In contrast, no significant correlation was found in the baseline condition ($r = .06$, 95% CI [−.13, .24], $p = .56$, $N = 111$) (Figure 9).

These findings suggest that in tool-enabled conditions (Masking and Filtering), participants' confidence was more closely aligned with the amount of supporting evidence they gathered, indicating a form of evidence-grounded confidence calibration. In contrast, participants in the baseline condition often reported high confidence regardless of how much evidence they collected, raising concerns about overconfidence or premature certainty in unguided audits.

## 5 Discussion

Our findings demonstrate that participants identified four patterns of gender bias in the model's behavior, including *stereotyped role assignments, the prioritization of gender over professional identity, biased interpretation of visual cues,* and *gendered language that reinforced traditional hierarchies.* The types of hypotheses participants generated were strongly shaped by the auditing tools they used. The image masking tool enabled fine-grained testing of visual and contextual cues, often revealing inconsistencies such as reversed gender roles when faces were obscured. The text filtering supported the detection of broader, systemic patterns, particularly asymmetries in how the model described men and women through language. In contrast, the baseline interface led to more instance-specific observations, typically grounded in clear but isolated examples of bias. These contrasts also shaped how confident participants felt in their hypotheses. We elaborate on each of these findings in this section.

### 5.1 User Audits Reveal How Gender Biases Interact in Image Captioning

Previous research on gender bias in image-captioning models has mainly relied on benchmark datasets and predefined metrics [1, 15, 16]. Although these studies provide useful characterizations of bias patterns, they often evaluate them in isolation and under controlled conditions, which limits our understanding of how multiple biases may interact or co-occur in real-world scenarios. In contrast, even within the delineated scope of occupational settings in the medical sector, our thematic analysis of user-driven audits revealed rich descriptions of many different biases and identified the precise circumstances under which each bias emerges. These audits, produced by 60 non-expert participants in 30 minutes after an introduction to the task and interface, captured all the bias pattern reported in the specialized literature on visual-language models. This suggests that, for this kind of representational harm, rich analyses of model behavior can emerge without extensive technical training.

Most of the themes address different forms of *stereotype reinforcement* [16, 49]. In particular, Theme 1 underlines stereotyped descriptions of roles and occupations, which are even reinforced through language (Theme 4). Thanks to the interactivity of the interface, participants could further investigate the factors responsible for stereotyping, such as the presence of multiple people with

mixed genders (T2.3). This led them to uncover issues with *pronoun resolution* [15], further interpreting them as the attribution of authority to men even against visual evidence. Participants surfaced many cases of *context-to-gender* and *gender-to-context bias* [31] that occur when identical visual elements trigger different role labels depending on perceived gender. This concerned both visual cues such as clothing (T3.1) or background (T1.2), and the resulting language (T4.2). Our thematic analysis surfaced clear cases of *gender–activity binding* [1] that emerges when the model infers a role from gender rather than from the actual activity depicted (Theme 2). Participants were critical of how issues of *binary gender framing* [42] and *misgendering* [15] affected such binding and the interpretation of visual cues. With the Masking tool, they were able to conduct fine-grained experiments regarding the effect of masking visual cues on the detection of gender and its downstream repercussions on stereotyping (Themes 2 and 3). With the filtering tool, participants' increased focus on the language modality surfaced many cases of *Markedness* [45] – the addition of gender markers to otherwise neutral nouns. This was evident in Theme 4, especially regarding their interaction with *gender–activity binding* regarding the vocabulary used in the description of roles and appearance (T4.2 and T4.3).

Participants' analyses therefore revealed how multiple bias mechanisms often interacted, reflecting a layered, compounding process embedded in model behavior. Many also reflected on how these patterns mirrored broader entangled societal hierarchies and embedded cultural narratives about gender and professional identity. These results underline the promise of user-driven audits grounded in the sensemaking framework to deliver diverse and high-quality reports of model behavior in specific settings. Although our dataset centered on gendered professional roles in medical imagery, the sensemaking-based design of the auditing interface, methods, and theoretical framing is not tied to this domain. The same workflow can be instantiated with other vision–language models and datasets organized around different forms of representational bias.

## 5.2 Interactive Tools Shape Sensemaking and Confidence

Across all conditions, participants engaged in the full arc of sensemaking: spotting harmful behaviors, forming hypotheses, collecting examples, and synthesizing broader patterns. These conditions shared a design grounded in the sensemaking framework as proposed by Cabrera et al. [4]. Our design shared several features with their platform AIFinnity, including the ability to browse data, create groups of evidence and maps evidence to hypotheses. While Cabrera et al. [4] examined sensemaking in a model comparison task with data scientists, our study applied the framework to a user-driven auditing task focused on gender bias, conducted by non-experts. Despite this shift in task and population, participants produced rich analyses that collectively captured patterns of bias previously identified in a fragmented manner across earlier studies [1, 16, 31, 45], demonstrating the breadth of insights enabled by a sensemaking-driven design.

However, our findings also show that the interface conditions significantly shaped the granularity of these observations and steered participants toward different layers of gender bias by focusing on visual features, linguistic framing, or their interaction. In this sense, auditing tools helped surface specific bias patterns but also constrained inquiry: filtering users rarely questioned why certain phrases recurred, while masking users did not compare caption phrasings. This illustrates how scaffolding steers audits direction by shaping what becomes visible and actionable.

In addition, the actions supported by the audit tools (Masking and Filtering) also influenced participants' sense of confidence in their audits. Filtering enabled users to surface repeated patterns at scale, which often reinforced certainty in their claims; masking, by contrast, invited more cautious interpretation due to the ambiguity of visual changes. Confidence is particularly relevant for user-driven auditing tasks, where insights are not externally verified but must be persuasive to other stakeholders. Thus, interface affordances not only shaped the sensemaking process by guiding users toward particular bias patterns, as also observed by DeVos et al. [10], but also affected how confident participants felt in articulating their conclusions and carrying out the task.

Although our analysis focuses on audit outcomes rather than process logs, differences across conditions suggest that each tool made specific sensemaking stages more salient, consistent with our design rationale in Section 3.3. In the masking condition, generating alternative versions of the same image seemed to serve as counterfactual probes, helping participants revisit the *Instances & Outputs* stage to refine hypotheses grounded in visual cues. In the filtering condition, keyword-based aggregation seemed to enable comparisons across grouped captions, supporting the identification of recurring patterns in gendered phrasing and occupational labels, which could made the *Schemas* stage more salient. The observed correlation between confidence ratings and number of evidence supports this, indicating that participants grounded their claims in aggregated patterns surfaced by the filtering tool.

Our work therefore contributes to the current endeavor of designing audit tools that facilitate discovering harms and communicating results, which has been shown to be less common than the tools that are used to evaluate models [8, 33].

## 5.3 Auditing as a Collective Sensemaking

Our findings show that participants often surfaced different, sometimes contradictory interpretations of the same behaviors. For example, some viewed the label "male doctor" as reinforcing dominant norms by implying that men naturally belong in certain professional roles. Others interpreted the reverse pattern, "female doctor," as marking women as exceptions in male-dominated professions. While centered on similar outputs, these interpretations reflected distinct hypotheses: one critiqued gender marking as biased toward a male default, while the other read it as signaling deviation from role expectations.

These perspectives highlight two insights. First, what counts as harm is not fixed, but filtered through the auditor's lens: participants' standpoints were shaped by factors such as demographic identity, cultural background, prior beliefs, and lived experience. This echoes Wieringa's emphasis on the role of the "forum" in accountability relationships: different observers require different accounts depending on their standpoint, and each forum, whether legal, professional, or civil society, asks different kinds of questions and makes different kinds of judgments [44].

Second, rather than treating one view as "correct," auditing is a space where multiple, equally valid perspectives can coexist. Prior research has shown that diverse perspectives can expand coverage while also deepening critique, prompting participants to reflect on their own assumptions as they challenge the model [6, 30]. Although our study focused on individual auditing, the interpretive variation we observed points to the latent potential of collective sensemaking. This resonates with Shen et al. [39]'s account of everyday algorithm auditing, where users often make sense of systems not alone but in conversation with others: discussing outputs, refining hypotheses, and negotiating what counts as harm. Similarly, there is a growing body of work on distributed or crowd sensemaking [12, 20], which explores how schemas and hypotheses from multiple people can be aggregated and reused [4]. From this perspective, the goal is not to assume unbiased auditors, but to design processes and tools that make different standpoints visible and comparable, so that disagreements can be examined and discussed collectively.

Future auditing tools should therefore move beyond supporting only individual reflection. They should enable collaborative sensemaking by allowing users to share schemas, compare hypothesis variations, and collectively broaden the space of critique. To support this, tools must actively scaffold multi-perspective engagement by connecting users across demographic and experiential backgrounds, and by surfacing disagreement not as noise, but as a necessary condition for accountability.

### 5.4 Limitations and Future Directions

While this study offers empirical insights into how non-expert users audit algorithmic behavior, it suffers several limitations. First, participants audited a fixed set of 80 images pre-selected for gender bias in occupational contexts, which oriented sensemaking toward occupational stereotypes while excluding other forms of harm, such as intersectional dynamics of race, age, or ability. Expanding the tasks or enabling user-defined datasets could surface more diverse biases. Second, although Prolific provided participants from an international audience and various sociodemographic background, our sample lacked intentional inclusion of marginalized communities, which likely constrained the perspectives and interpretive strategies observed. Recruiting participants with diverse lived experiences would both enrich coverage and test whether auditing tools can accommodate multiple cultural and experiential standpoints. Additionally, although we collected Ambivalent Sexism Inventory (ASI) to ensure that pre-existing attitudes did not differ across conditions, our analysis centered on interface design and sensemaking outcomes. Future studies could examine how specific factors such as sexist beliefs, feminist identification, or professional background, may influence how users interpret harm and engage with different auditing tools.

Moreover, our study focused on participants' final audit outcomes rather than tracking how they moved through different stages of sensemaking, which limits our ability to analyze how tool affordances shaped these trajectories over time. Future work could employ fine-grained interaction logging or think-aloud protocols to examine how users navigate sensemaking as a dynamic, multi-stage process.

Finally, the study was conducted in a controlled, individual setting, whereas many people encounter AI systems in everyday situations such as autocomplete, content feeds, or generative assistants. Future systems should embed auditing into these workflows through lightweight prompts, collaborative spaces, or contextual annotations, enabling situated and collective audits. This raises open questions for future work: how many auditors are needed to surface a broad range of harms, and how can divergent interpretations be collectively negotiated? Concepts like data saturation from qualitative research may help estimate when additional audits stop revealing new insights. Alternatively, collaborative auditing could offer ways to resolve disagreement and assess overall coverage. Addressing these questions calls for technical, social, and epistemic infrastructures that sustain auditing as an ongoing, reflective practice.

## 6 Conclusion

This work investigated how interface design choices driven by sensemaking processes impact user-driven algorithm auditing. We studied how non-experts engaged with an image captioning model to identify and reason about profession-related gender bias across three interfaces: a Baseline for open-ended exploration of image dataset and captions, a Masking Tool for manipulating visual inputs, and a Filtering Tool for querying the dataset based on words in the captions. Our findings show that interface design shaped what participants noticed, how they framed bias, and how they supported their hypotheses. The Masking Tool enabled fine-grained, counterfactual testing of visual cues and context, while the Filtering Tool revealed broader asymmetries in the model's use of gendered language. Across conditions, participants surfaced intersecting patterns of gender bias and interpreted model behavior as a complex interplay between visual and linguistic signals.

Future work should expand this approach to other domains and modalities, support collaborative and longitudinal auditing practices, and embed sensemaking-driven tools into everyday ML interaction. Designing for user-led interpretation will be essential for fostering more accountable and transparent AI systems.

## References

[1] Ali Abdollahi, Mahdi Ghaznavi, Mohammad Reza Karimi Nejad, Arash Mari Oriyad, Reza Abbasi, Ali Salesi, Melika Behjati, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2024. Gabinsight: Exploring gender-activity binding bias in vision-language models. *arXiv preprint arXiv:2407.21001* (2024).

[2] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction* 5, CSCW1 (2021), 1–34.

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

[4] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M Drucker. 2023. What did my AI learn? How data scientists make sense of model behavior. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–27.

[5] Souti Chattopadhyay, Zixuan Feng, Emily Arteaga, Audrey Au, Gonzalo Ramos, Titus Barik, and Anita Sarma. 2023. Make It Make Sense! Understanding and Facilitating Sensemaking in Computational Notebooks. *arXiv preprint arXiv:2312.11431* (2023).

[6] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[7] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[8] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[9] Wesley Hanwen Deng, Claire Wang, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *arXiv preprint arXiv:2501.01397* (2025).

[10] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[11] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 62–71.

[12] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 247–256.

[13] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 39–53.

[14] Peter Glick and Susan T Fiske. 2018. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social cognition*. Routledge, 116–160.

[15] Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems* 36 (2023), 63687–63723.

[16] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*. 771–787.

[17] A Hern. 2020. Twitter apologises for" racist" image-cropping algorithm', Guardian, 21 September.

[18] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 3819–3828.

[19] Kimon Kieslich, Nicholas Diakopoulos, and Natali Helberger. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment. *AI and Ethics* (2024), 1–23.

[20] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 999–1010.

[21] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.

[22] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[24] Rena Li, Sara Kingsley, Chelsea Fan, Proteeti Sinha, Nora Wai, Jaimie Lee, Hong Shen, Motahhare Eslami, and Jason Hong. 2023. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[25] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. 2024. "They only care to show us the wheelchair": Disability representation in text-to-image AI models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.

[27] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[28] Abhishek Mandal, Suzanne Little, and Susan Leavy. 2023. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 416–424.

[29] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[30] Behnoosh Mohammadzadeh, Jules Françoise, Michèle Gouiffès, and Baptiste Caramiaux. 2024. Studying collaborative interactive machine teaching in image classification. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 195–208.

[31] Yuta Nakashima, Yusuke Hirota, Yankun Wu, and Noa Garcia. 2023. Societal bias in vision-and-language datasets and models. *NIHON GAZO GAKKAISHI (Journal of the Imaging Society of Japan)* 62, 6 (2023), 599–609.

[32] Felicia Ng, Jina Suh, and Gonzalo Ramos. 2020. Understanding and supporting knowledge decomposition for machine teaching. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1183–1194.

[33] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *arXiv preprint arXiv:2402.17861* (2024).

[34] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[35] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32 (2020), 6363–6381.

[36] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating online research through sensemaking integration and machine research support. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 608–618.

[37] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 913–926.

[38] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[39] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[40] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[41] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[42] Eddie L Ungless, Björn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. *arXiv preprint arXiv:2305.17072* (2023).

[43] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 324–335.

[44] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 1–18.

[45] Robert Wolfe and Aylin Caliskan. 2022. Markedness in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1269–1279.

[46] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. 2021. RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–26.

[47] Yisong Xiao, Aishan Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Xianglong Liu, and Dacheng Tao. 2024. Genderbias-\emph {VL}: Benchmarking gender bias in vision language models via counterfactual probing. *arXiv preprint arXiv:2407.00600* (2024).

[48] Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

[49] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[50] Kankan Zhou, Yibin LAI, and Jing Jiang. 2022. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics.