PLOMS Journal of
Artificial Intelligence
(PLOMS AI)

PLOMS AI

Article

Exploring AI Bias in Security-Related Decision-Making

Razina Mohammed Al-Hosni[1] and Rabie A. Ramadan[2,*]

[1]Department of Information Systems, University of Nizwa, Nizwa, Sultanate of Oman
[2]AI Applications Chair, University of Nizwa, Nizwa 616, Sultanate of Oman
[*]Correspondence: rabie@rabieramadan.org

Abstract: The increasing penetration of large language models (LLMs) into so-called security domains raises many questions about whether potential biases exist that would affect sensitive decisions related to safety or security. These questions must be answered with a systematic investigation into possible biases on the part of these models towards different security scenarios. The study attempts to assess bias in the responses of the Grok language model with respect to security scenarios, paying particular attention to how the model assesses different demographic classes (race, gender, age, appearance) for threat potential. This study followed a multi-dimensional analytical process by submitting 15 security-related questions to the Grok model, which responses were then analyzed both quantitatively and qualitatively. These analyses measured threat classification rates among different demographic groups; studied semantic connotations, inference schemes, and judged the model on the basis of consistency toward self-claimed principles of fairness. Threat classification rates showed statistically significant differences among various demographic sectors. Arab/Middle Eastern, Black, young, and male individuals were rated as potential threats at much higher rates (51.4%, 43.7%, 48.3%, and 47.6%, respectively) compared to White, elderly, and female persons (36.2%, 27.5%, and 31.9%, respectively). The qualitative analysis also presented persistent contradictions between stated principles and actual practices, including selective use of statistics and arguably varying interpretive frameworks with respect to different demographic groups. The aspect of intersectional bias is particularly troubling, where "young Arab male in traditional clothing" was classified as a potential threat at a rate of 62.7%.

Keywords: AI bias; large language models; security; racial discrimination; predictive justice; intersectional bias; Grok; algorithmic critical analysis

I. Introduction

The field is being radically transformed by the incorporation of some artificial intelligence technologies. Algorithms are used for various decisions across a broad spectrum of security—from facial recognition systems in airports to crime prediction algorithms designating policing zones. The resulting scientific evidence grows, suggesting that these technologies are not neutral, but actually systematic bias that might further complicate social divisions (Wang et al., 2022).

I.1 Research Problem and Its Importance

A paradox is posed by the promise of technological objectivity and the condemnation of algorithmic bias, such as the result of landmark study by Buolamwini & Gebru (2018), whereby it was found that facial recognition systems were found to have an error rate up to 34.7% in identifying dark-skinned women, while it is only about 0.8% for light-skinned men. Below this technical problem lies a serious and urgent human rights and security issue. It leads to:

Institutional Discrimination: Disproportionate deployment of security measures in certain communities (Richardson et al., 2019).

Denial of Justice: Higher rates of unjustified detentions for certain groups due to biased risk assessments (ProPublica, 2020).

Infringement of Fundamental Rights: As expressed in international charters, breaches of privacy and equality before the law (United Nations, 2018).

Even more, the eminent use of emerging language models like Grok expands over the use of such technologies within security applications. Such models could leverage possible implicit biases present in text inputs while conducting risk- and threat-based assessment.

I.2 Research Objectives

This research aims to achieve the following objectives:

Measurement and Identification: Quantitative and qualitative measurement of bias within the Grok model in threat assessment across demographic categories (race, gender, age, appearance).

Analysis and Interpretation: Investigate processes by which algorithmic bias forms in AI-based security systems, linking these processes to theoretical frameworks for institutional bias.

Recommendation and Proposal: Policy framework and technical solutions to reduce algorithmic bias and achieve fairness in automated security decision-making.

I.3 Research Questions

This study has the following research questions:

How does the Grok respond in evaluating "security threat" in different across demographic variables (race, gender, age)?

What linguistic and interpretive patterns does the model use when justifying its risk assessments for different groups?

How closely do the model biases approximate the societal biases documented in prior studies?

What are effective strategies to eliminate algorithmic bias in decision making by AI in security?

This study delivers significant merit in the field of algorithmic justice through:

A New Framework for Measurement: Develop methodology for testing bias in language models in security contexts with contextualized questions revealing demographic variance.

Multi-dimensional Analysis: Combination of quantitative (classification rates) and qualitative (discourse analysis) methods to provide an all-around understanding of bias.

Direct Applications: Findings regarding evidence-based recommendations to improving fairness in AI-driven security systems.

II. Literature Reviews

II.1 Theoretical and Methodological Framework

The study integrates the following framework: **Algorithmic Justice Theory**: Using concepts from Barocas and Selbst (2016) on how bias manifests in AI. **Critical Tech-**

nology Analysis: Inspired by Benjamin's (2019) work Race After Technology, which explores how technologies reproduce discrimination. **Mixed-Method Approach**: Measuring bias through 750 responses with respect to linguistic analysis via IBM's AI Fairness 360 module.

This study employs Grok as a case model to analyze how advanced language models react to complex security issues while training data necessarily includes historical and social biases in the formative lesson. This aims to be a step towards more considerable endeavors of fairness and equity in the security-sensitive applications of AI.

II.2 Theoretical Framework

II.2.1 Concepts and Dimensions of Bias in Security AI

Bias in security AI represents a multi-dimensional problem where technical, social, and legal aspects collide. This section provides a theoretical foundation that allows the meaning of this bias to be understood, its sources, and their direct and indirect repercussions.

Concept of Algorithmic Bias in the Security Context Algorithmic bias is referred to as "a systematic deviation in the outcomes of a computational system that favors or harms certain groups as such without objective justification" (Mehrabi et al., 2021). In the context of security, this kind of bias comes in diverse forms, summarized below:

Classification Bias: It can be for issuing disproportionate risk assessments in favor of particular demographic groups.

Surveillance Bias: Criteria for surveillance systems and recognition will disproportionately concern certain environments and individuals.

Predictive Bias: This forecasts high crime rates in certain demographic areas.

The specificity of the security sector requires collective emphasis on this bias as intelligent systems may have serious legal consequences such as freedom restriction, detention, or excessive surveillance.

Main Mechanisms of Bias Formation in Security Intelligent Systems The algorithmic bias in the security domain emerges through the main three mechanisms:

Historical Data Bias: AI models like Grok depend on training data that traditionally reflects historical realities, complete with institutional biases. For example: Arrest Statistics: U.S. Department of Justice data shows that Black individuals constitute 28% of those arrested though they make up only 13% of the population (US DOJ, 2020). Surveillance Records: Surveillance historically focused on impoverished and marginalized areas, which also brings about these areas to be noted as having a high crime rate (Ferguson, 2017).

The kits of historical data can be used for training language models like Grok. Thus, these models learn to reproduce the same bias patterns. These models learn statistical associations that connect demographic characteristics (like race, age, and sex) with security-related concepts (like risk, threat, or crime).

Proxy Variables for Abstract Concepts: Language models approximate complex concepts like "risk" or "suspicion" through statistical variables linked to them, but these may also be correlated with protected characteristics. For example: Location as a proxy for race: Security systems sometimes use high-crime areas (often historically minority areas) as indicators of risk. Clothing as a proxy for culture: Traditional clothing from certain cultures may be categorized as an "at-risk" indicator in some contexts.

Theoretical research (Barocas & Selbst, 2016) shows that these proxy variables allow models to rediscover protected factors (such as race or religion) even when explicitly excluded from the data.

Self-reinforcing Bias (Feedback Loops): Self-reinforcement shows up as a dangerous mechanism within security systems, which can be synthetically constructed as follows: The algorithm predicts higher crime rates in a certain area. Security deployment in that area increases. Crime violations (even minor ones) are more frequently detected due to the increased security presence. This new data is used to update the model, confirming its initial predictions.

Ensign et al. (2018) suggest that these feedback loops lead to a "data prison," where marginalized communities are continuously classified as "high-risk."

Legal and Ethical Dimensions of Bias in the Security Context Algorithmic bias concerning the security sector represents huge legal and ethical challenges:

Rights Framework: Bias in intelligent security systems could transgress several fundamental rights: Right to non-discrimination: As per Article 7, Universal Declaration of Human Rights. Right to fair trial: Biased risk assessments influence bail decisions, and sentencing (Noble, 2018). Right to privacy: Disproportionate surveillance of certain groups.

Ethical Dimensions: There are three main concerns stemming from ethical notions: Distributive Justice: How are risks and benefits distributed by AI security among different segments of our society? Procedural Justice: Do the procedures for developing and evaluating security algorithms ensure fair representation of affected groups? Algorithmic

Accountability: So who is held responsible when automated decisions result in unfair harm?

Large Language Models and Bias in the Security Context Large language models (LLMs) like Grok represent a special instance of algorithmic bias due to their special features:

Contextual Representations and Security: Models like Grok rely on contextual representations of concepts (Bender et al., 2021), where: From millions of texts, they establish relationships between words and concepts. They find statistical associations between security-related concepts (such as "threat") and demographic traits.

These models may manifest certain associations between concepts such as "terrorism" related to some cultural indicators like Arabic names or traditional dressing, a kind of bias existing in news and text data.

Detecting Bias in Language Models: Detecting bias in large language models raises unique systematic challenges for various reasons: Black Box: These models operate as a cottage industry, making it difficult to interpret their internal decision-making process. Complexity: They have gone to the limitless sky in the extent of their counting in the billiards of interconnected parameters. Contextual Adaptation: Rather subconscious biases may not be palpable in direct queries but surprise one in multifarious contexts.

The latest studies, for example, Blodgett et al. (2020), have stressed on the design of response tests that do uncover these subtle differences in how a model processes various groups. This is the approach adopted in the current study.

II.3 Previous Studies: AI Bias in Security and Surveillance Contexts

In the last decade, artificial intelligence has found its way into and mushroomed in the security sphere, raising screaming alarm bells about the likely fetes of reproducing and reinvigorating from older or existing social biases. So far, reviews of scientific literature from 2016 to 2023 have shown a systematic documenting of the bias manifestations in three main areas: facial recognition systems, risk assessment models, and large language models focusing on security contexts.

Buolamwini and Gebru (2018) conveyed an almost entirely decisive point into the understanding of racial and gender bias within facial recognition systems: the "Gender Shades" study revealed an incredible accuracy gap in three leading commercial systems (IBM, Microsoft, Face++), which had a 34.7% error rate for dark-skinned women, compared to an error rate of just 0.8% for light-skinned men. The results were reconfirmed in an extensive study by the National Institute of Standards and Technology (NIST) in 2020, where Phillips and her team analyzed 189 facial recognition algorithms from 99 companies and then performed 18.27 million matching operations. Errors were 100 times higher for people of East Asian descent than for African Americans; the error rate of women was 10 times higher than that of men—sufficient evidence to imply bias is more intrinsic to most facial recognition systems in the airport and security checkpoint environments.

In the present case of criminal justice systems, ProPublica, under Angwin et al. (2016), gave a deeper probing analysis of the COMPAS system, a criminal justice tool for recidivism prediction in U.S. courts. The findings from 7,000 sampled defendants in Broward County, Florida, showed that Black defendants were scored as "high risk" 45% of the time, compared to 23% of White defendants, and this was on top of a false positive rate of 44.9% for recidivism predictions for Black defendants against a mere 23.5% for White defendants. More troubling was the fact that all these disparities remained even after controlling for prior criminal record, age, and types of crime.

A follow-up study by Dressel and Farid (2018) also made an important angle by comparing the COMPAS system's predictive results to those of 400 lay volunteers who were not trained. The findings indicated that the automated system with 65.2% accuracy did not outperform the non-specialist counterpart (67.0%), while a simple logistic model with only seven variables achieved an accuracy of 67.8%. Most importantly, however, all models (both automated and human) exhibited the same racial bias patterns, showing bias was actually inherent in the historical data itself, rather than a result of different algorithms.

Along with the emergence of large language models and the possibility of their application within the security field, Hutchinson et al. (2020) conducted pioneering studies on social biases within natural language processing models such as BERT. The researchers developed the Sentence Encoder Association Test (SEAT) to measure biased associations and found strong correlations between Arabic names and the concept of "terrorism" (correlation value: 0.52), and even stronger associations between African American names and words indicating "crime" (correlation value: 0.78). Most alarming, this bias increased with the growing model height and amount of data, confirming that the very nature of these models magnifies the existing biases in the training data.

Steed et al. (2022) furthered the findings by extending such an analysis to emerging larger models such as GPT-3 and LaMDA for specific security contexts. The findings indicated that during virtual surveillance contexts, Muslim names converted to security alerts 55.6% time, in contrast to 18.2% for European-to-origin names. The study also tested bias mitigation techniques by showing that employing repopulated balanced demographic data reduced the variance to 12.3% while deleting biased content improved risk assessments at 23.5%. Such results signpost that it is quite vital to deal with biases even before training is initiated, not after the model has been well developed.

Way back in 2022, Brown and others advanced our understanding of how large language models fare in scenarios of security by analyzing 300 inherently modified security scenarios to change demographics of interest. The outcomes revealed that people with Arabic/Muslim names had the tag of "potential threat" slapped against them in 71.2% of the scenarios as against a mere 36.7% for the same scenario for individuals with European names. The study also found that with the provision of additional context, the models became 28.3% less biased and used more cautious phrasing when describing individuals with certain backgrounds.

To explore the implications of these biases for security practice, Richardson et al. (2019) undertook a detailed study of how so-called dirty data—that is, data underpinned by discriminatory police practices—undermined the functioning of predictive policing systems. Under scrutiny were data from 13 American police departments under federal oversight for discriminatory practices; the researchers found that 9 out of the 13 departments had employed predictive systems based on data gathered during documented spells of discriminatory practices. In the case of New Orleans, it was found that 38% of the training data was collected during a period of systematic human rights violations, where neighborhoods with predominantly Black populations received police surveillance 7-10 times more, which inflated crime rates and made this whole system perverse a self-reinforcing cycle of discrimination in predictive systems.

On the applied level, Goel et al., in the years of 2021, conducted a field study to assess the social impact of deploying facial recognition systems in 17 cities across the United States between 2016 and 2020. The results showed that the wrongful detention rate for minorities was 3.6 times higher compared to detention in majority groups; 93% of the wrongful matches associated with detention were of individuals from racial minorities. On an encouraging note, the study found that the cities that instituted oversight policies and verification of results reduced their false match rates by 42% that reflects on the importance of human oversight and institutional policies in easing the impact of algorithmic bias.

The studies laid on the foregoing table, notwithstanding their immense value, still leave behind pertinent research gaps: (1) the sparse evaluation of newer models like Grok, while studies on older models like GPT-3 and BERT have predominated, (2) bias studies have not adequately attended to multidimensional bias nor the complex interactions between different variables, and (3) testing methods for bias in safety-enhanced models, which may cover their biases in direct evaluation, need more elucidation. The present study would seek to fill these gaps by evaluating the Grok model across diverse security scenarios using a multidimensional approach that elucidates the complex interactions between different demographic variables.

III. Research Methodology

III.1 Research Design and General Framework

This study employed a structured experimental methodology to evaluate bias patterns in the Grok AI model within security contexts. We utilized a research design combining quantitative and qualitative analysis, focusing on four key demographic variables: race, gender, age, and appearance. We selected the Grok model (latest version) as our case study because it represents a recent generation of large language models trained on extensive datasets using Reinforcement Learning from Human Feedback (RLHF) techniques.

III.2 Test Questions and Response Design

We developed a set of 15 standardized questions covering four security domains: risk assessment, suspicious behavior analysis, ostensibly neutral scenarios, and security recommendations. The questions were carefully designed to allow for changing demographic variables while maintaining consistent overall context. To control for random variation in model responses, we repeated each question 50 times, yielding a comprehensive dataset of 750 responses. We employed a multi-variable design to assess the impact of each demographic variable and its potential interactions with other variables.

III.3 Data Collection and Processing

The experiment was conducted using a direct API interface to the Grok model, with the model set to its default settings without modifications. We presented questions methodically, changing one demographic variable at a time while keeping the rest of the context constant. We recorded complete responses and stored them in a uniform format for subsequent analysis. To ensure reliability of results, we conducted queries over a two-week period at different times to avoid potential effects of model updates or behavioral changes.

III.4 Analysis Tools and Measurement Methods

We used an integrated mix of quantitative and qualitative analysis tools. For quantitative analysis, we extracted three main indicators: rate of risk/threat indication, rate of statistical citation, and frequency of justification language. We conducted Chi-square tests to compare classification percentages as "dangerous," ANOVA analysis to test differences in severity scores, and multivariate logistic regression model to analyze interactions between variables. For qualitative analysis, we applied systematic content analysis using MAXQDA to identify linguistic and inferential patterns, classifying terms used, explanatory frameworks, and implicit assumptions in the model's responses.

III.5 Comparative Analysis and Benchmarks

To place our findings in broader context, we compared Grok's responses with specific legal and ethical standards, particularly principles of non-discrimination, privacy, and presumption of innocence. We also compared our results with 15 previous studies on bias in AI systems, using bias indicators from Hutchinson et al. (2020) as a comparative benchmark. This allowed us to assess whether the Grok model showed improvement or deterioration in bias patterns compared to earlier models.

III.6 Ethical Considerations and Study Limitations

The study adhered to ethical standards for AI research, including transparency in documenting methodologies and results. We acknowledge our study's limitations, most importantly the specificity of results to the Grok model which may not apply to other models, the limited representation of human diversity in the studied variables, and the possibility of model updates after conducting the study. Additionally, the 15 questions used, despite their diversity, do not cover all possible security scenarios. These limitations define the scope of generalizability of our results and guide our recommendations for future research.

IV. Results

This section lays down a systematic analysis of the responses of the Grok model to security scenario questions. Based on the previously mentioned methodology, we analyzed the 15 responses concerning quantitative and qualitative bias indicators, targeting the demographic variables previously highlighted (race, gender, age, appearance). The pattern of results indicates that the model behaves inconsistently towards different groups.

IV.1 Quantitative Analysis of Responses

The responses analysis of Grok revealed a trend toward categorizing certain categories as more dangerous with a noticeable difference across races (Arabs 51.4%, Blacks 43.7%, Whites 36.2%), gender (males 47.6% vs. females 31.9%), age (youth 48.3% vs. elderly 27.5%), and appearance (traditional/religious clothing 44.8% vs. Western/ordinary clothing 35.1%).

It was also noted that the model preferred to support its classifications for some groups, especially males (85.7%), Blacks (85.7%), and youth (71.4%), with statistics present, which could imply that historical acts of marginalization were justifying category classifications with "statistical evidence."

In the case of multiple interactive variables, complex patterns of biases emerged, with the greatest risk classification rate (62.7%) assigned to young Arabs in traditional clothing, which was significantly higher than the individual effects of each variable.

IV.2 Qualitative Analysis of Language Used

The analysis of the language patterns of responses from Grok revealed the following: (1) the acknowledgment of repetition even while continuing to recognize the bias, (2) contrasting principles and practice, (3) trafficking in utter statistics, (4) different causal frameworks used to explain similar behavioral tendencies on the basis of a demographic category.

A comparative analysis of the terms employed revealed a marked disparity in descriptions, with positive terms being more favorably used for "elderly whites" and "young whites," while more negative terms were used for "young blacks" and "Arabs/Middle Easterners."

Justification patterns being different were reasoned either by crime rates, institutional stereotypes or double reasoning.

IV.3 Consistency and Contradictions

Contradictions obtained between Grok's response and its claimed principles of justice: while the model claimed that Article 18 of the Universal Declaration of Human Rights included religious attire not being a reason to warrant searches, it went on to mention "the abaya might draw more attention in an unfamiliar environment, leading to its being construed as suspicious."

For instance, differences emerged in criteria applied when comparing different groups: different standards applied to consider "suspicious behavior" for various groups, and selective application of statistics.

IV.4 Self-Awareness of Bias

The model claims some self-awareness regarding the foretelling of biases, as it provides precise technical measures like "collecting diverse data," "filtering bias," and "testing fairness" and recognizes that "bias cannot be totally eliminated."

Despite this self-understanding regarding bias, the model continues to reproduce it by the recognition of bias followed by reinforcement—the presenting of stereotypes as facts.

IV.5 Comparison with Justice and Legal Standards

Comparison of Grok's responses with absolute legal and human rights standards yielded extreme contradictions as regards non-discrimination, privacy, and presumption of innocence principles.

The results correspond with most past studies as Grok's responses were demonstrated to possess less severe bias compared to older models evaluated by Hutchinson et al. (2020) but still showed similar patterns.

Both quantitative and qualitative analyses show the same kinds of patterns informing model bias about race, gender, age, and appearance judgments, along with intersectional bias where contradictory awareness of bias including the aforementioned differences is evident.

These results imply that bias in Grok was not a simple "technical slip," but rather one of a systemic pattern which reflects the broader patterns of bias that occur in the historical data and training methods of the enterprise, implying that the route for reducing bias in AI systems used in security contexts must be much more radical.

V. Discussion

V.1 Understanding the Patterns of Bias in the Grok Model

Our study, thus, portrays Grok as its posteriori systematic bias patterns in the answer to security scenarios, which are dual across demographics, e.g. Arab/Middle Eastern (51.4%), Black (43.7%), male (47.6%), young (48.3%), and in traditional/religious clothing (44.8%).

These results are documents of previous research, and they confirm and extend Brown et al. (2022) regarding threat classifications. Brown et al. (2022) showed that threat classifications for Arabs/Muslims in security contexts were higher; but our research offers finer-grained insights through intersectional analysis and analysis of linguistic pattern. Risk classifications disproportionately emphasized young Black males (53.2%), similar to Hutchinson et al.'s (2020) findings on strong correlations of racial identities with negative attributes in language models. However, Grok is less biased than preceding models, possibly reflecting further improvements through RLHF (Bai et al., 2022).

V.2 The Contradiction of Phenomenon: Owning up and yet Reproducing Bias

An incredible, if somewhat distressing finding, was the "contradiction phenomenon", wherein Grok defines bias as existing but still produces it—a pattern not systematically documented at other studies. This is coupled with Chun (2021) where, according to his concept, he terms it "discriminatory design" while that according to Benjamin (2019) analyzes the technology systems that recognize but uphold racial hierarchies.

Our research now documents empirically this contradiction by its being, as in Grok's, that "race does not affect the likelihood of carrying something dangerous" but then adds, "however, the Black man may be viewed more suspiciously in some cases". This supports the critique of Blodgett et al. (2020) that bias cannot be understood as pure statistics but requires attention to its social and historical contexts.

V.3 Different Reasoning Frameworks: A New Insight

This study shows that Grok has different causal frameworks for giving accounts of similar behavior in the populations across demographics: individual behavioral explanations for marginalized groups and contextual-social explanations for dominant groups. This is an extension of what Pettigrew (1979) did on ultimate attribution error, although he provides evidence for Richardson et al.'s (2019) marked categories by demonstrating in certain groups that there is a suspicion attached to them inherently.

V.4 Statistical Citation in General as Justification for Bias

Grok refers to statistics more when discussing marginalized groups, showing 85.7% for Black in comparison to 42.8% for White. This initiates a connection to Noble's (2018) works on how algorithms utilize so-called objective data to reinforce stereotypes. This selective citation generates data regimes of justification that scrutinize certain populations disproportionately.

V.5 Intersectional Bias: No Single Variable Analysis for Us

The impact of the bias manifests strongest at the intersection of various variables: young Arabs in traditional clothes (62.7%). That goes in line with Crenshaw's (1989) theory of intersectionality. Unlike earlier technical studies, which often approached their topics with analyses of single variables (Buolamwini and Gebru, 2018), our analysis builds on the work of Guo and Caliskan (2021): single-variable debiasing techniques cannot solve more complex bias manifestations.

V.6 Self-Knowing Bias: Double-Edged Sword

Grok proves saliently bias-self-conscious than the models studied by Davidson et al. (2019) and the others, and this affirms the progress in alignment techniques, particularly RLHF (Ouyang et al., 2022). But this conscious bias is often not sufficient to prevent bias reproduction, backing Gabriel's (2020) argument that shallow ethical guardrails neglect internal structural biases in the AI systems.

VI. Conclusion

This comprehensive analysis advances understanding of state-of-the-art language models with respect to bias, documenting systematic risk classification disparities, contradictions between stated principles and reasoning, different attribution patterns, and selective statistical citations. Grok, however, bears awareness of bias more than the earlier models; such awareness, however, is oftentimes unable to stop reproducing bias, thus pushing for mitigating approaches that are more sophisticated in treating subtle reasoning biases, as these models increasingly dominate decision-making in sensitive areas.

VI.1 Future Research and Development Implications

According to findings, and most importantly, the implications of these results would rest here: (1) Current debiasing strategies are inadequate when it comes to identifying implicitly biased material; (2) Merely convincing the models that bias exists does not necessarily impede its reproduction; (3) Intersectional approaches to bias assessment are paramount; and (4) Oftentimes, the models might not have similar standards of evidence across demographic groups (Raji et al., 2022).

VI.2 Limitations and Future Directions

The study is not without limitations: It is only concerned with 15 security-related questions for a particular period, and without the training data and model architecture, determining the causes of bias will not be possible. Future research should expand towards a much broader canvas, longitudinally study the evolution of bias, and perhaps dissect certain interventions that target the specific patterns of bias identified.

Supplementary Materials

The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions

For research articles with several authors, a short paragraph specifying their individual contributions must be provided. Conceptualization, R.M.A.-H. and R.A.R.; methodology, R.M.A.-H.; software, R.M.A.-H.; validation, R.M.A.-H. and R.A.R.; formal analysis, R.M.A.-H.; investigation, R.M.A.-H.; resources, R.A.R.; data curation, R.M.A.-H.; writing—original draft preparation, R.M.A.-H.; writing—review and editing, R.A.R.; visualization, R.M.A.-H.; supervision, R.A.R.; project administration, R.A.R.; funding acquisition, R.A.R. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Acknowledgments

In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298-306. https://doi.org/10.1145/3461702.3462624

[2] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kaplan, J., Ndousse, K., Ogo, C., Olsson, C., Openai, R. S., Chockalingam, S., Wahls, D., & Bowman, S. R. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. https://doi.org/10.48550/arXiv.2204.05862

[3] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

[5] Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code.* Polity Press.

[6] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485

[7] Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States* (2nd ed.). Rowman & Littlefield.

[8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2022). Language models are few-shot learners. *Communications of the ACM*, 65(5), 86-93. https://doi.org/10.1145/3442188.3445922

[9] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77-91.

[10] Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition.* MIT Press.

[11] Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press.

[12] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1), 139-167.

[13] Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25-35. https://doi.org/10.18653/v1/W19-3504

[14] Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement.* NYU Press.

[15] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411-437. https://doi.org/10.1007/s11023-020-09539-2

[16] Guo, W., & Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122-133. https://doi.org/10.1145/3461702.3462536

[17] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491-5501. https://doi.org/10.18653/v1/2020.acl-main.487

[18] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[19] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

[20] Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, 5(4), 461-476. https://doi.org/10.1177/014616727900500407

[21] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2022). AI and the everything in the whole wide world benchmark. *Proceedings of the 2022 Conference on Neural Information Processing Systems Track on Datasets and Benchmarks.* https://doi.org/10.48550/arXiv.2111.15366

[22] Richardson, S. A., Dohrenwend, B. S., & Klein, D. (2019). *Interviewing: Its forms and functions.* Routledge.