Uncovering Discriminatory Behavior in LLM-Based CV Screening Systems with Composite Statistical Metamorphic Relations

Tetyana Turiy , **Simon Speth** and **Alexander Pretschner**
Technical University of Munich
{tetyana.turiy, simon.speth, alexander.pretschner}@tum.de

Abstract

Artificial intelligence (AI) applications are prone to reproducing existing patterns of marginalization, bias, and discrimination. The adoption of automated decision-making tools, particularly those based on large language models (LLMs), necessitates the detection and analysis of their discriminatory behavior to ensure compliance with anti-discrimination regulations. This work advances metamorphic testing for bias detection, aiming to (1) uncover discriminatory behavior in AI tools and (2) standardize the metamorphic test creation approach, rooting it in the legal and academic research on discrimination. To fill in these gaps, we introduce a metamorphic test creation workflow and evaluate it on the LLM-based CV screening use case. Our method yields a metamorphic relation (MR) catalogue of 8 atomic and 6 composite MRs, legally justified by Article 21 (1) of the Charter of Fundamental Rights of the European Union and sociologically motivated by the intersectionality theory. The resulting metamorphic test suite uncovers discriminatory failures for all 6 evaluated LLMs, totaling 1,272 defects across 7,800 test cases. Both the new atomic MRs and the composite MRs detect additional discriminatory defects for all LLMs. Composite MRs consistently uncovered more defects than their atomic counterparts, with failure rates increasing by on average 9% for Llama-3.1 8B and 25% for Llama-3.2 3B.

1 Introduction

Context. Discrimination based on conscious or unconscious bias is a major issue in our society. Members of marginalized communities experience difficulty accessing services due to unfavorable treatment by decision makers, who might be guided by prejudice, fear, or lack of knowledge about the impact of their actions [Tasioulas, 2019]. With the increasing use of autonomous decision-making tools reproducing such bias, discriminatory behavior might (unintentionally) be automated and replicated at scale [O'Neil, 2016; Broussard, 2023]. Some of the examples are a discontinued Amazon hiring software, which rated resumes with the word "women's" in it worse [Dastin, 2018], the UK immigration decision-support system with a category "for speedy boarding for white people from the most favored countries" [McDonald, 2020], and a car insurance system, where foreign-born drivers were systematically charged higher premiums [Fabris *et al.*, 2021].

Problem Domain. In this work, we rely on software testing, particularly metamorphic testing (MT), to detect discriminatory behavior. MT enables the creation of automated test oracles to test how the system's behavior changes across multiple inputs in relation to each other [Chen *et al.*, 2018; Chen *et al.*, 1998]. This is an especially promising approach for bias detection, as it allows for complex test cases comparing the system's treatment of members of the *marginalized groups* and members of the *dominant groups*[1]. MT is acknowledged as a powerful bias detection tool in the natural language processing (NLP) domain, especially for the tasks of sentiment analysis [Khoo *et al.*, 2023; Asyrofi *et al.*, 2022; Ma *et al.*, 2020], natural language inference [Li *et al.*, 2024], machine translation [Sun *et al.*, 2024], and question answering [Hyun *et al.*, 2024]. However, (1) protected attributes with respect to which fairness is measured are inconsistent across the works, and (2) the rationale for the choice of both the attributes and identity groups is rarely provided. Additionally, (3) the interaction between multiple attributes and their combined impact on the system is not sufficiently explored.

Solution. To bridge the gap between anti-discrimination legislation and software engineering efforts to detect discriminatory behavior, we propose a test creation workflow that derives metamorphic relations (MRs) based on legally protected attributes, for example, a list of protected attributes in Article 21 of the Charter of Fundamental Rights of the European Union (CFR)[2]. To ground MT in real-world discrimination research, we propose relying on sociological correspondence studies to elicit failure hypotheses, i.e., theories about which

Replication package: https://doi.org/10.5281/zenodo.20004501

[1] Marginalized groups are systematically disadvantaged in society, for example, through limited access to employment, whereas dominant groups hold disproportionate power, resources, and influence, for example, through preferential treatment when seeking employment.

[2] Article 21 (1) of the Charter of Fundamental Rights of the European Union: "Any discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited".

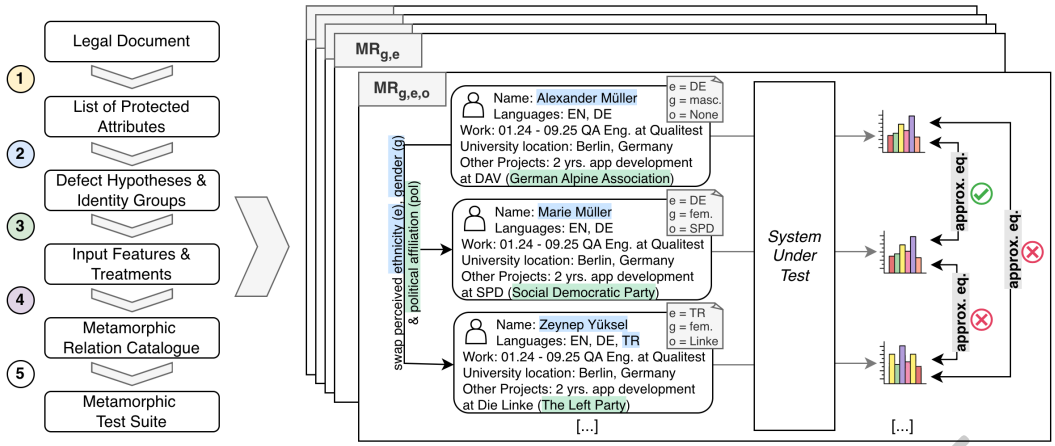

Figure 1: Approach overview with test creation workflow on the left and its evaluation with CV screening use case on the right.

inputs may cause a discrepancy between the expected and observed outcomes. Correspondence studies are a gold standard technique for measuring discrimination in hiring, housing, or consumer market research, where a researcher modifies one or more attributes (such as race, age, gender, etc.) in an application, while controlling for others (income, education, etc.) and then measures the difference in decision-makers' responses [Gaddis, 2018; Verhaeghe, 2022]. Additionally, motivated by intersectionality theory [Crenshaw, 2013], we advocate expanding fairness testing to include intersectional identities, accounting for the unique experiences that occur at the intersection of multiple axes of privilege and marginalization. In this work, we refer to MRs, where only one protected attribute is modified between the source and follow-up inputs as *atomic*, and MRs where multiple attributes are modified as *composite* [Liu *et al.*, 2012; Qiu *et al.*, 2022].

Contribution. The following artifacts are the main contributions of this work: (1) a test creation workflow to evaluate automated decision-making systems, which results in a metamorphic test suite where each MR is legally and sociologically justified, and (2) its evaluation on the curriculum vitae (CV) screening use case with large language models (LLMs).

2 Background

Defect-Based Testing. Defect is an umbrella term for the concepts of a fault, error, and failure. A failure is a deviation of expected and actual system behavior, an error is a deviation of expected and actual program state, and a fault is the actual cause of the deviation [Laprie, 1992]. Defect-based testing methods posit that test selection is effective when each test case is derived from a corresponding defect hypothesis [Pretschner, 2015; Speth *et al.*, 2025; Speth *et al.*, 2026].

In this work, we advocate for defect-based testing with a fault hypothesis that artificial intelligence (AI) models may reproduce discrimination patterns, similar to those internalized by humans, due to training on historical data that is prone to sampling bias and prejudice in labels, as well as human-imposed assumptions during modeling and data engineering [O'Neil, 2016; Ferrer *et al.*, 2021; Broussard, 2023].

Metamorphic Testing. Metamorphic testing (MT) is a test case generation technique, centered around the concept of instantiating MRs into multiple metamorphic test cases (MTCs), where MRs are "necessary properties of the target function or algorithm in relation to multiple inputs and their expected outputs" [Chen *et al.*, 2018]. In an MTC, a system under test (SUT) is applied at least twice to at least one source input and at least one follow-up input, which is generated from the source input by applying a metamorphic transformation function. The source and follow-up outputs are the outputs of the SUT applied to the source and follow-up inputs, respectively. To determine whether the MTC passes or fails, one checks whether a predefined relation holds between the outputs.

By defining a test verdict through a relation between the source and follow-up outputs, MT solves the oracle problem without the need for manually created test outcomes. Because of this, we consider MT especially well-suited in the context of bias detection, because differential treatment can be defined as a violation of an *approximately equal* relation between the outcomes for a member of a marginalized group and outcomes for a member of the dominant group.

3 Related Work

MT for Bias Testing. The limited availability of labeled test datasets, as well as the risk that LLMs might ingest those datasets during training, has led MT to become a popular tool for evaluating LLMs. A survey by Cho *et al.* identifies 44 works and extracts 191 unique MRs, from which 36 MRs were implemented and executed on 4 NLP-related tasks [Cho *et al.*, 2025]. Some of the 191 MRs refer to bias testing by substituting names, pronouns, countries, and occupations.

We performed a literature review using Scopus, IEEE Xplore, and the ACM Library with the following search string: *"metamorphic testing" AND ("bias" OR "discrimination" OR "fairness")*. After removing duplicates, we are left with ten relevant papers. The MRs they implemented, and their mapping to the corresponding CFR protected attributes are summarized in Table 1. No protected attribute is featured consistently in every work. Moreover, the rationale for the choice of attributes is either not provided or only refers to the "broad so-

Attributes in CFR	# works	References
Race/ethnic origin	8	Gir25, Red25, Sri25, Hyu24, Li24, Sun24, Kho23, Asy22
Gender	7	Rom25, Hyu24, Li24, Sun24, Kho23, Asy22, Ma20
Social origin	7	Gir25, Rom25, Red25, Sri25, Li24, Kho23, Asy22
Religious belief	4	Gir25, Rom25, Red25, Sri25
Language	3	Gir25, Red25, Sri25
Political opinion	3	Gir25, Red25, Sri25
Age	2	Hyu24, Li24
Sexual orientation	2	Rom25, Hyu24

Table 1: Overview of protected attributes evaluated in related work on MT for discriminatory behavior detection. For the lack of clear mapping to protected attributes in the CFR, some attributes (e.g., marital status, physical appearance) are not depicted. When operationalized as MRs, these are further referred to as the *base* set MR_B.

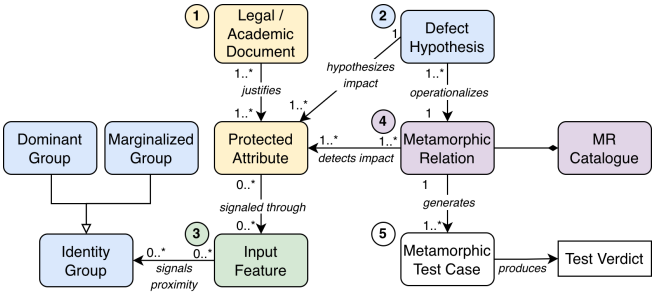

Figure 2: Metamodel of our proposed approach to identify defect-based MRs for discriminatory bias detection with steps from Figure 1.

cial relevance and their prevalence in related work" [Romero-Arjona *et al.*, 2025]. This work is also the only one to explain the choice of identity groups included in testing. Only three works have explored intersectional MRs [Giramata *et al.*, 2025; Reddy *et al.*, 2025; Srinivasan and Abdel, 2025]. All the identified works above rely on a template-driven approach, where attributes (e.g., 'Asian', 'young', 'low-income') are directly inserted or swapped; in contrast, we use correspondence studies to elicit more complex metamorphic transformation functions.

Contrary to comparing the LLM outputs based on semantic similarity, as it is common in NLP-related MT [De Curtò and De Zarzà, 2025; Tu *et al.*, 2021], we use statistical MT [Guderlei and Mayer, 2007; Rehman and Izurieta, 2021] to account for the stochastic nature of LLM outputs, while focusing on quantifiable outcomes such as CV rating.

Auditing LLM-based Decision-Making. Signaling gender and race/ethnicity through names is a common practice in sociology that was also applied to LLM evaluation. Nghiem *et al.* and Salinas *et al.* evaluated LLMs performance on multiple decision-making tasks, including hiring and salary recommendations, by comparing responses to short template-based prompts with names of a job seeker varying by racial and gendered perception [Nghiem *et al.*, 2024; Salinas *et al.*, 2024]. Similar to our approach Armstrong *et al.* and Gaebler *et al.* used real CVs with name substitutions to detect potential discrimination in hiring decisions [Armstrong *et al.*, 2024; Gaebler *et al.*, 2024]. Seshadri *et al.* uses not only names to signal gendered and racial categories, but also extracurricular activities (e.g., "Black Student Union" or "Women in Engineering") to strengthen demographic signals [Seshadri *et al.*, 2025]. We, in contrast, propose incorporating more treatments from correspondence studies into bias testing, allowing us to evaluate the SUT's behavior towards more identity groups.

4 Evaluation Method

This Section introduces our proposed approach, aiming to standardize MT for bias detection. Figure 2 illustrates a metamodel of our 5-step approach, which begins with the identification of protected attributes in legal documents and results in test verdicts from the execution of the metamorphic test suite.

First (step ①), an appropriate set of protected attributes is selected as the foundation for MR development. Protected attributes are personal traits legally safeguarded against discrimination in areas like employment, housing, etc. We propose to start with the list of protected attributes in Article 21 of the CFR[3]. Naturally, MRs should be guided, but not limited to, protected attributes defined by law. We recommend expanding the list of protected attributes beyond the strict letter of the law, based on ethical considerations, as it might take time for something that is societally understood to be discriminatory and harmful to be declared so legally.

Next (step ②), defect hypotheses are derived based on domain knowledge. Use case specific identity groups that the SUT may display prejudice towards are selected. The foundational fault (i.e., reason for failure) hypothesis of this work is that AI models may reproduce the same discriminatory patterns internalized by humans. Therefore, we recommend referring to sociological research to identify discriminatory behavior (failure hypotheses) to test.

In step ③, we identify features of the input that can signal proximity to the previously identified identity groups. Usually, protected attributes are not mentioned explicitly, but signaled through (a combination of) proxy attributes. For example, gender and ethnicity might be perceived through first and last names. In our experience, correspondence studies [Verhaeghe, 2022; Gaddis, 2018] are a good starting point for identifying treatments (i.e., modifications of inputs) that can then be implemented as transformation functions.

In step ④, relations between source and follow-up outputs are defined, so that MRs can be implemented. In this work, we limit ourselves to the *approximately equal* relation, in which the system is expected to remain consistent in its ranking of members of marginalized and dominant groups. Alternatively, this can be implemented as a *greater than or equal* for outputs for the marginalized group compared to the dominant group (i.e., marginalized-group subjects are treated no worse than dominant-group subjects). It is also possible to derive concrete failure hypotheses for the concrete identity groups (e.g.,

[3]Alternatively, Equality Act 2010 in the UK, the Canadian Human Rights Act, various anti-discrimination law in the USA (e.g., Civil Rights Act of 1964, Americans with Disabilities Act of 1990, Pregnancy Discrimination Act, etc.), German General Equal Treatment Act, Article 3 Abs. 3 of the German Constitution.

candidates who do not disclose any religious affiliation are preferred to those who do according to secularization theory [Wright *et al.*, 2013]). However, it is worth noting that there exist multiple theoretical models of discrimination [Fibbi *et al.*, 2021] that, if instantiated for a particular set of identity groups in a particular context, might result in competing failure hypotheses. Therefore, we choose to create approximately equal relations and further evaluate whether any of the theoretical models apply based on test logs.

Lastly (step ⑤), the MR catalogue is implemented, and each MR is automatically instantiated into multiple MTCs. Then, all MTCs are executed and statistically evaluated. In this work, we use the GeMTest metamorphic testing framework [Speth and Pretschner, 2025] to implement transformation functions and MRs and automatically execute the test suite.

5 Experiments

In this Section, the test creation workflow introduced in Section 4 is evaluated by applying it to an LLM-based CV screening procedure motivated by the increasing automation and use of AI tools in various stages of the recruitment process [Drage and Mackereth, 2022; Bhatnagar *et al.*, 2025; Lo *et al.*, 2025].

5.1 Research Questions

The following experiments target such research questions:
RQ1: Is our proposed test creation workflow effective in producing a test suite for defect-based bias detection?
RQ2: How many additional defects can be discovered by the set of new MRs compared to the set of base MRs?
RQ3: How many additional defects can be discovered by the set of composite MRs compared to the set of atomic MRs?

5.2 Experiment Setup

We set out to evaluate the robustness of a range of LLMs employed for CV screening. To mimic the experience of a recruiter, who might have a limited experience with LLMs and prompt engineering, but also to simplify the testing setup, we employ a relatively simple process of CV screening: in a chat session, we request an LLM to provide a grade from 0 to 100 for a text-based CV and a short job description. We focus on the similarity of the grades a SUT assigns to the candidates' CVs, as opposed to, for example, the semantic similarity of textual responses, so that we can target quantifiable outcomes for the candidates in our evaluation.

Dataset. From a dataset of synthesized semi-structured CVs[4], we sample 20 CVs for each of the five job titles with the most entries in the dataset, namely database administrator, backend developer, deep learning engineer, embedded systems engineer, and quality assurance engineer. We add a summary to the CVs and use the job search website *indeed.com* to manually select a description and five required skills for the jobs to provide to the LLM for guidance in candidate evaluation.

Systems Under Test. We evaluate six LLMs to assess their ability for unbiased CV evaluation. To ensure reproducibility of our experiments, we choose not to use any commercially available APIs, such as those by OpenAI, as we do not know all the pre- and post-processing steps performed, nor which particular model version is deployed at any given moment, making reproducibility impossible. We use the GPT4All [Anand *et al.*, 2023] wrapper to deploy all LLMs and request the models to grade a CV in a chat session (as opposed to text completion mode). Our model selection was restricted by resource and checkpoint availability and resulted in the following 6 SUTs: Llama 3.2 1B, Llama 3.2 3B, Phi 3 Mini 4B, Nous Hermes 2 7B, Llama 3.1 8B, and Llama-3-8B DeepSeek distillation.

MR	Protected Attribute(s)	Treatment	Perceived Identity Groups	Study
MR_g	Gender	Name	Feminine, Masculine	Lan21
MR_e	Ethnic or., Race, Color, Lang., National min.	Name, Lang.	German, Polish, Russian, Turkish, Syrian	Lan21
MR_p	Parenthood	Volunteering	Parent, Control	Cor07
MR_r	Religion/Belief	Volunteering	Evangelical, Catholic, Muslim, Control	Pie13
MR_o	Political opinion	Volunteering	CDU, SPD, Grüne, Linke	Gif15
MR_d	Disability	Disclosure in CV summary	Blindness, Deafness, Autism, Spinal cord injury, Control	Bae16
MR_q	Sexual orientation	Volunteering	LGBTQIA+ identity, Control	Wei15
MR_s	Social origin	Hobby, Volun.	"Lower", "Upper" class	Riv16

Table 2: Overview of the MRs implemented in the CV screening bias detection test suite and treatments to signal proximity to different identity groups are motivated by correspondence studies in labor market discrimination.

5.3 Results of RQ1

In this Section, we describe the test suite derivation process resulting from applying the test creation workflow. Table 2 summarizes MRs implemented for the CV screening use case.

Protected Attributes. Starting with the list of protected attributes from Article 21 (1) of the CFR fixed in step ①, we focus in step ② on the overview of 90 correspondence studies on hiring discrimination conducted between 2005 and 2016 [Baert, 2018] to identify relevant marginalized and dominant groups and derive failure hypotheses.

Failure Hypotheses. The base failure hypothesis for each MR is that members of different identity groups may face different outcomes, even if only the values of protected attributes without job performance influence are modified. Some concrete failure hypotheses rooted in theoretical models of discrimination are evaluated based on test logs and discussed in Section 6, but not implemented in the test suite.

Identity Groups. Identity groups, towards whom the SUT's behavior is evaluated, are selected based on the German cultural context due to the authors' background, and have been selected based on German demography: four largest immigrants groups in 2023[5]; three largest religious groups in 2024[6], political parties scoring over 5% in the 2025 Bundestag election and represented in the 21st German Bundestag[7] not classified as a "confirmed right-wing extremist endeavor" by the German Federal Office for the Protection of the Constitution (BfV).

[4]https://huggingface.co/datasets/datasetmaster/resumes

[5]https://www.statistischebibliothek.de/mir/receive/DEHeft_mods_00164830

[6]https://fowid.de/meldung/religionszugehoerigkeiten-2024

[7]https://www.bundestag.de/dokumente/textarchiv/2025/kw11-bundeswahlausschuss-1056272

Model	MR_g	$MR_{g,e}$	MR_p	$MR_{p,g}$	$MR_{p,g,e}$	MR_r	$MR_{r,g}$	$MR_{r,g,e}$	MR_o	$MR_{o,g}$	$MR_{o,g,e}$	MR_d	$MR_{d,g}$	$MR_{d,g,e}$	MR_q	$MR_{q,g}$	$MR_{q,g,e}$	MR_s	$MR_{s,g}$	$MR_{s,g,e}$	fr_{SUT}
Llama 3.2 1B	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.007
Llama 3.2 3B	0.04	0.20	0.02	0.07	**0.28**	0.13	**0.28**	**0.46**	0.10	0.19	**0.37**	**0.46**	**0.51**	**0.67**	0.04	0.09	0.25	0.05	0.11	0.26	0.253
Phi 3 Mini 4B	**0.10**	**0.33**	**0.11**	0.09	0.26	0.11	0.15	0.32	**0.18**	0.15	0.36	0.43	0.47	0.57	**0.06**	0.07	0.24	**0.12**	0.12	0.27	**0.258**
Nous Hermes 2 7B	0.08	0.25	0.06	**0.11**	0.25	**0.16**	0.23	0.33	0.14	**0.20**	0.35	0.42	0.47	0.60	0.03	**0.14**	**0.27**	0.06	**0.19**	**0.34**	0.251
Llama 3.1 8B	0.04	0.12	0.02	0.05	0.13	0.09	0.12	0.19	0.06	0.09	0.14	0.32	0.33	0.41	0.03	0.02	0.10	0.09	0.10	0.18	0.145
DeepSeek Distill 8B	0.03	0.03	0.00	0.01	0.07	0.01	0.02	0.10	0.05	0.01	0.12	0.06	0.11	0.23	0.00	0.00	0.07	0.02	0.00	0.10	0.066
$\overline{fr_{MR}}$	0.05	0.16	0.04	0.06	0.17	0.09	0.13	0.23	0.09	0.11	0.22	0.29	0.32	0.41	0.03	0.05	0.16	0.06	0.09	0.19	

Table 3: Failure rate per MR with Bonferroni correction, where MR_x represents an atomic MR, $MR_{x,y}$ and $MR_{x,y,z}$ represent 2-composite and 3-composite MRs with two and three protected attributes modified simultaneously. The highest failure rate in each column is marked in **bold**. fr_{SUT} refers to the failure rate of the SUT on the entire test suite and $\overline{fr_{MR}}$ denotes the average failure rate per MR.

Metamorphic Relations. In step ②, we combine several protected attributes (ethnic origin, race, color, language, national minority) into one *ethnic origin* MR. We choose to add *parenthood* as a separate MR [Correll *et al.*, 2007] and not to implement the *age* and *genetic features* MRs for the lack of consensus on how to model different ages [Baert *et al.*, 2016] and a general lack of correspondence studies on discrimination based on genetic features [Baert, 2018].

Metamorphic Transformation. In step ③, we identify common treatments used in correspondence studies to use as metamorphic transformation functions. A candidate's name can be used to signal gender identity and race/ethnicity [Lancee, 2021]. Volunteering and work experience with various organizations affiliated with political or politicized causes can be used to signal support or belonging [Correll *et al.*, 2007; Pierné, 2013; Gift and Gift, 2015; Weichselbaumer, 2015; Rivera and Tilcsik, 2016]. Certain information, such as disability, can be disclosed in the CV summary [Baert, 2016].

In addition to every atomic MR in Table 2, a composite MR [Liu *et al.*, 2012; Qiu *et al.*, 2022] is implemented to investigate the effects of the intersectional axis of marginalization and privilege of ethnic/national origin and gender categories. We use the shorthand notation $MR_{p,g,e}$ for $MR_p \circ MR_g \circ MR_e$, which denotes the composition of MR_p, MR_g, and MR_e, where the protected attributes 'political opinion', 'gender', and 'ethnicity' are modified simultaneously.

Statistical Relation Function. To account for the variability of the LLM outputs, the CV grading request is performed 30 times, each in a new session with context reset, a temperature of 0.7, and a top-p of 0.4, in order to get a big enough sample of the grade distribution for statistical testing. A test case fails if, for at least one pair of outputs in an MTC, the null hypothesis that the grades are sampled from the same distribution can be rejected based on a 2-sample Kolmogorov-Smirnov test with an α of 0.01 to capture differences in both mean and variance. To account for multiple pairwise tests and reduce the possibility of false positives, we use the Bonferroni correction.

Implementation. In step ⑤, the catalogue of MRs is implemented using the domain-specific language provided by the GeMTest metamorphic testing framework [Speth and Pretschner, 2025], which enables the automated creation and execution of MTCs as pytest test cases. A replication package containing the source code is provided.

Metrics. To judge the effectiveness of the catalogue of MRs for bias detection, we use failure rate (fr) as the main evaluation metric, which can be calculated by dividing the number of failed MTCs by the total number of MTCs.

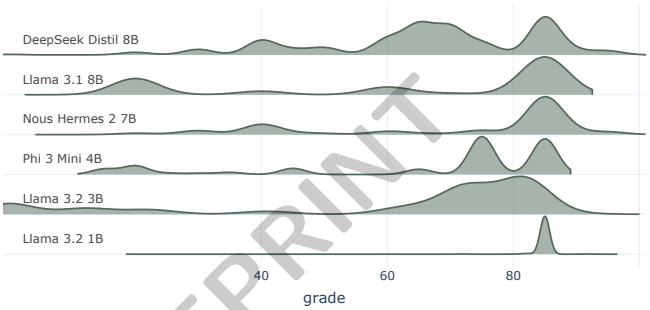

Figure 3: Kernel density estimation of grades assigned by the LLMs to CVs of varying quality (60% good fit, 40% bad fit).

Robustness Evaluation. Table 3 provides detailed results of the test suite execution with Bonferroni correction for multiple pairwise tests within an MR (see the supplementary material for the results before the correction and a more relaxed Holm-Bonferroni correction). The most robust model, i.e., the model with the lowest fr_{SUT}, is Llama 3.2 1B with a fr_{SUT} of 0.7% and only 9 failures uncovered. The least robust model is Phi 3 Mini 4B with a fr_{SUT} of 25.8% corresponding to 339 failures. Each MR uncovers failures for all SUTs, apart from the Llama 3.2 1B model, where only 5 out of 8 atomic and no composite MRs uncover failures, and DeepSeek Distill 8B, where 2 atomic MRs, MR_p and MR_q, do not reveal any defects. Thus, the catalogue of MRs resulting from our proposed test creation workflow is effective at detecting discriminatory behavior.

Discussion of RQ1 Findings. We run another experiment to evaluate the discriminatory power[8] of LLMs's as CV screeners, for which we sample CVs that are either a good or a bad fit for the job description. 60% of sampled candidates had previously held a job with a similar title, and 40% did not. Figure 3 provides a visualization of CV grade distributions per model, which demonstrates that Llama 3.2 1B is unfit to differentiate between good and bad candidates, supposedly because of the relatively small size for an LLM. We speculate that if Llama 3.2 1B cannot reject bad candidates, it is also not sensitive to input modifications to recognize belonging to different identity groups, which is the target of our MRs. This explains the 0.0 failure rate for most MRs. Still, according to our definition,

[8]"Discrimination" here means the ability to differentiate between fitting and unfitting CVs, not disparate treatment of some candidates.

Model name	MR_B	MR_N	$MR_{N\setminus d}$	MR_A	$MR_{A\setminus d}$	MR_C	$MR_{C\setminus\{d,g,e\}}$
Llama 3.2 1B	6	3	1	9	7	0	0
Llama 3.2 3B	52	48	2	100	54	229	162
Phi 3 Mini 4B	80	54	11	134	91	202	146
Nous Hermes 2 7B	64	48	6	112	70	214	154
Llama 3.1 8B	38	34	2	72	40	114	73
DeepSeek Distill 8B	11	6	0	17	11	69	46
Total failed MTCs	251	193	22	444	273	828	581
Total executed MTCs	3000	1200	600	4200	3600	3600	3000

Table 4: Failed MTCs per set of base atomic MRs MR_B, new atomic MRs MR_N, all atomic MRs MR_A, and compositional MRs MR_C. Due to the high failure rate of the disability MRs, we also present results for the new and composite sets excluding MR_d and $MR_{d,g,e}$.

the Llama 3.2 1B model is the most robust among the SUTs.

RQ1 Summary: By applying the test creation workflow, we derive an MR catalogue of 8 atomic and 6 composite MRs based on protected attributes from Article 21 (1) of the CFR and 7 correspondence studies on hiring discrimination. The resulting metamorphic test suite uncovers discriminatory failures for all evaluated LLMs.

5.4 Results of RQ2 and RQ3

Table 4 provides an overview of the defects detected by the four sets of MRs: (1) a base set of atomic MRs from prior work (MR_B) including sex/gender, social and ethnic origin, political opinion, religious belief, and sexual orientation; (2) a set of new atomic MRs proposed in this work (MR_N), including disability, and parenthood; (3) a set of atomic MRs (MR_A); and (4) a set of composite MRs (MR_C). We use the shorthand notation $MR_{C\setminus d}$ to denote the set $MR_C \setminus MR_d$.

Both the new atomic and the composite MRs detect additional discriminatory defects across all SUTs, with a total of 193 and 828 previously unknown failures detected with the new atomic MRs and composite MRs, respectively. Composite MRs consistently uncover more defects than their atomic counterparts. In Table 3, we see an increase in failure rate for almost all sets of MRs, where gradually one, two, and three protected attributes are modified. For example MR_r (religious affiliation), $MR_{r,g}$ (religious affiliation and perceived gender for the ethnic majority group), and $MR_{r,g,e}$ (composition of religious affiliation, perceived gender, and perceived ethnicity), increases for Phi-3 Mini 4B from 0.11 to 0.15 to 0.32 and for Llama-3.1 8B from 0.09 to 0.12 to 0.19. With each added axis of marginalization, new discriminatory defects are revealed.

RQ2 and RQ3 Summary: Both the new atomic and the composite MRs detect additional discriminatory defects for all SUTs. Composite MRs consistently uncover more defects than their atomic counterparts. The average increase in failure rate between an atomic and composite MR is 13%.

6 Discussion

We now discuss a selection of discriminatory behavioral patterns detected through analyzing the test logs.

In MR_r we observe a strong preference for candidates signaling affiliation with the Protestant Church, with a strong gen-

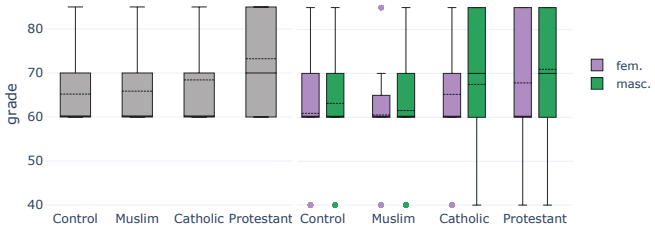

Figure 4: Distribution of CV grades assigned by the Llama 3.1 8B model in failed MTCs of MR_r (left) and $MR_{r,g}$ (right). In each box plot, a dashed line shows the mean and a solid line shows the median.

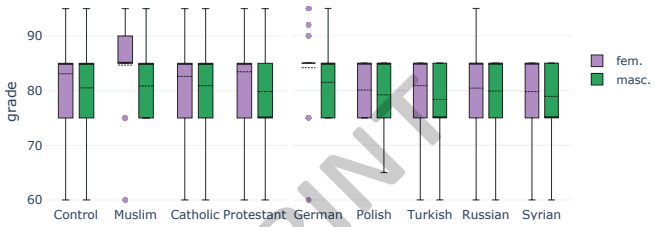

Figure 5: Distribution of CV grades assigned by the Nous Hermes 2 7B model in failed MTCs of $MR_{r,g}$ (left) and $MR_{r,g,e}$ evaluated only for candidates signaling affiliation with Islam (right).

dered component if evaluated at the intersection of religious affiliation and perceived gender in $MR_{r,g}$ for the Llama 3.1 8B model (see Figure 4). $MR_{r,g}$ also shows that the preference for Catholic candidates only extends to candidates perceived as men and has a limited positive effect for candidates perceived as women. Overall, candidates perceived as men receive a slightly higher average grade across all religious groups for Llama 3.1 8B (see Figure 4). However, the opposite effect is observed for Nous Hermes 2 7B with the strongest preference for women affiliated with Islam signaled through volunteering for a Muslim youth organization (see Figure 5). When evaluated at the intersection of gender and ethnicity, we see that women with names seen as German benefit from such volunteering more compared to candidates perceived to be from Muslim-majority countries, with candidates seen as Turkish Muslim men graded lowest on average.

If we compare the biases of the Llama-family models uncovered by gender/ethnicity $MR_{g,e}$, as visualized in Figure 6, the discriminatory behavior of the bigger model is more aligned with the biases consistent with the taste-based and contact theory of discrimination [Fibbi *et al.*, 2021]. We speculate that a larger model size correlates with more predictable human-like behavior learned from training data. Comparing the grades produced by Llama 3.1 8B, we observe a clear preference for candidates perceived to be German men; candidates with names perceived as Syrian are graded the lowest on average.

For all models, a strong effect of political affiliation is observed, but no consistent pattern across all models can be identified (see Figure 7). Notably, volunteering for the left-wing party Die Linke is negatively affecting the grades assigned by Phi-3 Mini 4B, having no political affiliation and volunteering for the Social Democrats (SPD) has a negative effect on Llama-3.1 8B. However, affiliation with SPD has a slight positive effect on grades assigned by the DeepSeek Distill

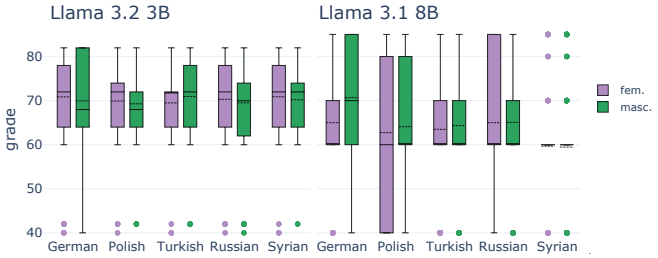

Figure 6: Distribution of CV grades assigned by Llama 3.2 3B (left) and Llama 3.1 8B (right) in failed MTCs from $MR_{g,e}$.

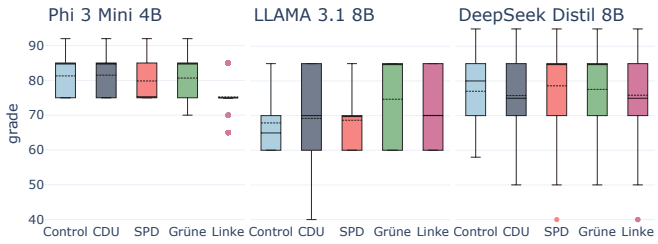

Figure 7: Distribution of CV grades assigned by Phi-3 Mini 4B, Llama-3.1 8B and DeepSeek Distill 8B models in failed MR_o MTCs.

8B model. All three models were in agreement in assigning slightly higher grades to candidates affiliated with the Greens.

7 Limitations

There are two levels of abstraction when discussing the limitations of our proposed approach: from an identity group to a treatment applied to the CV in the form of modifying one of the input features, and to the concrete values that signal belonging to one of the identity groups. Thus, while we follow the best practices established in labor market research on how to create CVs with strong signifiers of different identities, the treatments we apply might carry weak, mixed, or ambiguous signals, especially when processed by a program. In the following paragraphs, we provide examples of how these limitations may have influenced our results.

Weak Signal in Treatments. We opt into testing for gender-based discrimination only for binary gender signals from names, as research into which names are perceived as gender non-conforming in different cultural contexts is an open research area. We acknowledge that trans and gender non-conforming people face discrimination in hiring and harassment in the workplace [Taylor and Fasoli, 2022; Rosich, 2016], but the signal from names may be too weak for testing with our method and obfuscate rather than uncover potential harm.

Mixed Signal in Treatments. We observe a very high failure rate in the MR investigating disability, as well as a positive effect of disability disclosure, especially for autistic candidates. After reexamining the disclosure statements inspired by the work of Baert [Baert, 2016], we conclude that we add too much additional information, for example, "I benefit from regularity and structure, but this does not mean that I do not love a challenge" for autistic candidates. The added soft skills may contribute to this positive effect.

Ambiguous Signal in Treatments. Names of political parties are highly context-dependent. For example, Social Democrats and the Greens exist in many countries, varying in popularity and placement on the left-right political spectrum. Future work should explore concrete beliefs through membership in activist organizations, e.g., a workers' union.

Resource Limitations. Due to resource availability, the LLMs tested in this work are limited to models that require at most 16 GB VRAM. Thus, we cannot make statements about the discriminatory behavior of LLMs of greater size. However, we expect the bias-detection method to remain applicable.

Risk of Inconsistent Results. As much as this approach aims to increase reproducibility and transparency of the bias-focused test suite derivation, the test creation process is not deterministic. Firstly, the initial list of protected attributes depends on the choice of the legal document. For example, marriage is a protected attribute in the 2010 UK Equality Act, but not in the CFR. Secondly, we encourage testers to add attributes if it is reasonably justifiable, for example, through existing correspondence studies or discrimination theories.

8 Conclusions

This work proposes defect-based metamorphic testing for bias detection in AI tools, aiming to prevent the reproduction of systemic discrimination of marginalized communities by LLMs. Our main contribution is a test creation workflow that results in an MR catalogue, where every MR is traceable to a legal document and every metamorphic transformation function to a sociological correspondence study, justifying their necessity. The test creation workflow is evaluated on an LLM-based CV screening procedure to derive 8 atomic and 6 composite MRs, motivated by the intersectional theory. All MRs uncover failures for all SUTs, apart from the Llama 3.2 1B model, where the SUT was found unfit for the task. Thus, the MR catalogue resulting from our proposed test creation workflow is effective at detecting discriminatory behavior. In particular, the new atomic and composite MRs are a valuable contribution, as they detect additional discriminatory defects across all SUTs.

Composite MRs consistently uncover more defects than their atomic counterparts, as seen in the increase in failure rate of on average 9% and 25% for Llama-3.1 8B and Llama-3.2 3B, respectively. We find that some discriminatory patterns are undetectable with an atomic MR, but become more pronounced in a composite MR. We consider this the most important finding of this work and advise developers working on fair AI models to investigate behavior towards complex identities at the intersection of marginalization and privilege.

Future work should focus on evaluating LLMs that exceed the 16 GB of required VRAM and identifying additional protected attributes and MRs for a more robust bias assessment of AI-based software. Sources for these MRs could be, for example, expert interviews with legal scholars, anti-discrimination advocates, and people who experience discrimination.

Acknowledgments

The authors are grateful to Derui Zhu and Lena Gregor for their valuable feedback and helpful suggestions on earlier drafts of this paper.

References

[Anand *et al.*, 2023] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all, 2023.

[Armstrong *et al.*, 2024] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. The silicon ceiling: Auditing gpt's race and gender biases in hiring. In *EAAMO*. ACM, 2024.

[Asyrofi *et al.*, 2022] Muhammad Hilmi Asyrofi, Zhou Yang, Imam Nur Bani Yusuf, Hong Jin Kang, Ferdian Thung, and David Lo. Biasfinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *TSE*, 48(12):5087–5101, 2022.

[Baert *et al.*, 2016] Stijn Baert, Jennifer Norga, Yannick Thuy, and Marieke Van Hecke. Getting grey hairs in the labour market. an alternative experiment on age discrimination. *Journal of Economic Psychology*, 57:86–101, 2016.

[Baert, 2016] Stijn Baert. Wage subsidies and hiring chances for the disabled: some causal evidence. *The European Journal of Health Economics*, 17(1):71–86, 2016.

[Baert, 2018] Stijn Baert. *Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005*, pages 63–77. Springer, Cham, 2018.

[Bhatnagar *et al.*, 2025] Shatakshi Bhatnagar, Srijan Shetty, Nikhil Arora, Vishal Sachdev, and Aram Bahrini. Beyond traditional biases in ai hiring: Exposing the hidden systemic challenges in resume screening. In *SIEDS*, pages 280–285. IEEE, 2025.

[Broussard, 2023] Meredith Broussard. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press, Cambridge, Massachusetts, 2023.

[Chen *et al.*, 1998] Tsong Yueh Chen, Shing Chi Cheung, and Siu Ming Yiu. Metamorphic testing: A new approach for generating next test cases. Technical report, The Hong Kong University of Science and Technology, 1998.

[Chen *et al.*, 2018] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, TH Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *CSUR*, 51(1):1–27, 2018.

[Cho *et al.*, 2025] Steven Cho, Stefano Ruberto, and Valerio Terragni. Metamorphic testing of large language models for natural language processing. In *ICSME*, pages 174–186, 2025.

[Correll *et al.*, 2007] Shelley Joyce Correll, Stephen Benard, and In Paik. Getting a Job: Is There a Motherhood Penalty? *American Journal of Sociology*, 112(5):1297–1339, 2007.

[Crenshaw, 2013] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex. In *Feminist legal theories*, pages 23–51. Routledge, 2013.

[Dastin, 2018] Jeffrey Dastin. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. In: *Reuters*. https://reut.rs/2Po4ZJi, 2018. Accessed: 2025-06-08.

[De Curtò and De Zarzà, 2025] Joaquim De Curtò and Irene De Zarzà. Metamorphic testing for semantic invariance in large language models. *IEEE Access*, 13:214772–214791, 2025.

[Drage and Mackereth, 2022] Eleanor Drage and Kerry Mackereth. Does AI Debias Recruitment? Race, Gender, and AI's "Eradication of Difference". *Philosophy & Technology*, 35(89), 2022.

[Fabris *et al.*, 2021] Alessandro Fabris, Alan Mishler, Stefano Gottardi, Mattia Carletti, Matteo Daicampi, Gian Antonio Susto, and Gianmaria Silvello. Algorithmic audit of italian car insurance: Evidence of unfairness in access and pricing. In *AIES*, pages 458–468, 2021.

[Ferrer *et al.*, 2021] Xavier Ferrer, Tom Van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. Bias and discrimination in ai: a cross-disciplinary perspective. *TSM*, 40(2):72–80, 2021.

[Fibbi *et al.*, 2021] Rosita Fibbi, Arnfinn H. Midtbøen, and Patrick Simon. *Theories of Discrimination*, pages 21–41. Springer, Cham, 2021.

[Gaddis, 2018] Steven Michael Gaddis, editor. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer, Cham, 2018.

[Gaebler *et al.*, 2024] Johann D. Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. Auditing large language models for race & gender disparities: Implications for artificial intelligence-based hiring. *Behavioral Science & Policy*, 10(2):46–55, 2024.

[Gift and Gift, 2015] Karen Gift and Thomas Gift. Does Politics Influence Hiring? Evidence from a Randomized Experiment. *Political Behavior*, 37(3):653–675, 2015.

[Giramata *et al.*, 2025] Suavis Giramata, M. Srinivasan, V. N. Gudivada, and Upulee Kanewala. Efficient fairness testing in large language models: Prioritizing metamorphic relations for bias detection. In *AITest*, pages 191–200, 2025.

[Guderlei and Mayer, 2007] Ralph Guderlei and Johannes Mayer. Statistical metamorphic testing testing programs with random output by means of statistical hypothesis tests and metamorphic testing. In *QSIC*, pages 404–409, 2007.

[Hyun *et al.*, 2024] Sangwon Hyun, Mingyu Guo, and M. Ali Babar. Metal: Metamorphic testing framework for analyzing large-language model qualities. In *ICST*, pages 117–128, 2024.

[Khoo *et al.*, 2023] Lin Sze Khoo, Jia Qi Bay, Ming Lee Kimberly Yap, Mei Kuan Lim, Chun Yong Chong, Zhou Yang, and David Lo. Exploring and repairing gender fairness violations in word embedding-based sentiment analysis model through adversarial patches. In *SANER*, pages 651–662, 2023.

[Lancee, 2021] Bram Lancee. Ethnic discrimination in hiring: comparing groups across contexts. Results from a cross-national field experiment. *Journal of Ethnic and Migration Studies*, 47(6):1181–1200, 2021.

[Laprie, 1992] Jean-Claude Laprie. *Dependability: Basic Concepts and Terminology*, pages 3–245. Springer, 1992.

[Li *et al.*, 2024] Zhahao Li, Jinfu Chen, Haibo Chen, Leyang Xu, and Wuhao Guo. Detecting bias in llms' natural language inference using metamorphic testing. In *QRS-C*, pages 31–37, 2024.

[Liu *et al.*, 2012] Huai Liu, Xuan Liu, and Tsong Yueh Chen. A new method for constructing metamorphic relations. In *ICQS*, pages 59–68, 2012.

[Lo *et al.*, 2025] Frank P.-W. Lo, Jianing Qiu, Zeyu Wang, Haibao Yu, Yeming Chen, Gao Zhang, and Benny Lo. Ai hiring with llms: A context-aware and explainable multi-agent framework for resume screening. In *CVPR Workshops*, pages 4184–4193, 2025.

[Ma *et al.*, 2020] Pingchuan Ma, Shuai Wang, and Jin Liu. Metamorphic testing and certified mitigation of fairness violations in nlp models. In *IJCAI*, pages 458–465, 2020.

[McDonald, 2020] Henry McDonald. Home office to scrap 'racist algorithm' for uk visa applicants. In: *The Guardian*. https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants, 2020. Accessed: 2025-06-08.

[Nghiem *et al.*, 2024] Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé III. "You gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. In *EMNLP*, pages 7268–7287. Association for Computational Linguistics, 2024.

[O'Neil, 2016] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group, USA, 2016.

[Pierné, 2013] Guillaume Pierné. Hiring discrimination based on national origin and religious closeness: results from a field experiment in the Paris area. *IZA Journal of Labor Economics*, 2(1):4, 2013.

[Pretschner, 2015] Alexander Pretschner. Defect-based testing. *Dependable Software Systems Engineering*, 84:141–163, 2015.

[Qiu *et al.*, 2022] Kun Qiu, Zheng Zheng, Tsong Yueh Chen, and Pak-Lok Poon. Theoretical and empirical analyses of the effectiveness of metamorphic relation composition. *TSE*, 48(3):1001–1017, 2022.

[Reddy *et al.*, 2025] Harishwar Reddy, Madhusudan Srinivasan, and Upulee Kanewala. Metamorphic testing for fairness evaluation in large language models: Identifying intersectional bias in llama and gpt. In *SERA*, pages 239–246, 2025.

[Rehman and Izurieta, 2021] Faqeer ur Rehman and Clemente Izurieta. Statistical metamorphic testing of neural network based intrusion detection systems. In *CSR*, pages 20–26, 2021.

[Rivera and Tilcsik, 2016] Lauren A. Rivera and András Tilcsik. Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review*, 81(6):1097–1131, 2016.

[Romero-Arjona *et al.*, 2025] Miguel Romero-Arjona, José A. Parejo, Juan C. Alonso, Ana B. Sánchez, Aitor Arrieta, and Sergio Segura. AI-driven fairness testing of large language models: A preliminary study. In *SANER-C*, pages 25–32, 2025.

[Rosich, 2016] Gina Rosich. *Employment discrimination among people who are transgender or gender non-conforming: A mixed methods secondary data analysis*. PhD thesis, Fordham University, 2016.

[Salinas *et al.*, 2024] Alejandro Salinas, Amit Haim, and Julian Nyarko. What's in a Name? Auditing Large Language Models for Race and Gender Bias, 2024.

[Seshadri *et al.*, 2025] Preethi Seshadri, Hongyu Chen, Sameer Singh, and Seraphina Goldfarb-Tarrant. Small changes, large consequences: Analyzing the allocational fairness of LLMs in hiring contexts. In *NeurIPS*, 2025.

[Speth and Pretschner, 2025] Simon Speth and Alexander Pretschner. GeMTest: A general metamorphic testing framework. In *ICSE-Companion*, pages 41–44, 2025.

[Speth *et al.*, 2025] Simon Speth, Maximilian Trien, Dominik Kufer, and Alexander Pretschner. Safety-critical oracles for metamorphic testing of deep learning lidar point cloud object detectors. *OJ-ITS*, 6:95–108, 2025.

[Speth *et al.*, 2026] Simon Speth, Tobias Springer, Jannik Fischbach, and Alexander Pretschner. Traceable metamorphic test cases for robust safety-critical systems: A deep learning lidar object detector example. *STVR*, 36(3-4):1–21, 2026.

[Srinivasan and Abdel, 2025] Madhusudan Srinivasan and Jubril Abdel. Genfair: Systematic test generation for fairness fault detection in large language models. In *AITest*, pages 182–190, 2025.

[Sun *et al.*, 2024] Zeyu Sun, Zhenpeng Chen, Jie Zhang, and Dan Hao. Fairness testing of machine translation systems. *TOSEM*, 33(6), 2024.

[Tasioulas, 2019] John Tasioulas. First steps towards an ethics of robots and artificial intelligence. *Journal of Practical Ethics*, 7(1):49–83, 2019.

[Taylor and Fasoli, 2022] Edward Taylor and Fabio Fasoli. Who is most discriminated against? First impression and hiring decisions about non-binary and trans woman candidates. *Psicologia sociale*, 17(3):381–405, 2022.

[Tu *et al.*, 2021] Kaiyi Tu, Mingyue Jiang, and Zuohua Ding. A Metamorphic Testing Approach for Assessing Question Answering Systems. *Mathematics*, 9(7):726, 2021.

[Verhaeghe, 2022] Pieter-Paul Verhaeghe. Correspondence Studies. In Klaus F. Zimmermann, editor, *Handbook of Labor, Human Resources and Population Economics*, pages 1–19. Springer, Cham, 2022.

[Weichselbaumer, 2015] Doris Weichselbaumer. Testing for Discrimination against Lesbians of Different Marital Status: A Field Experiment. *Industrial Relations: A Journal of Economy and Society*, 54(1):131–161, 2015.

[Wright *et al.*, 2013] Bradley R. Entner Wright, Michael Wallace, John Bailey, and Allen Hyde. Religious affiliation and hiring discrimination in new england: A field experiment. *Research in Social Stratification and Mobility*, 34:111–126, 2013.