To be published in the Proc. of the 2nd International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications (FAIEMA 2024)

Testing software for non-discrimination: an updated and extended audit in the Italian car insurance domain

Marco Rondina, Antonio Vetrò, Riccardo Coppola, Oumaima Regragrui, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto and Juan Carlos De Martin

Abstract. Context. As software systems become more integrated into society's infrastructure, the responsibility of software professionals to ensure compliance with various non-functional requirements increases. These requirements include security, safety, privacy, and, increasingly, non-discrimination.

Motivation. Fairness in pricing algorithms grants equitable access to basic services without discriminating on the basis of protected attributes.

Method. We replicate a previous empirical study that used black box testing to audit pricing algorithms used by Italian car insurance companies, accessible through a popular online system. With respect to the previous study, we enlarged the number of tests and the number of demographic variables under analysis.

Results. Our work confirms and extends previous findings, highlighting the problematic permanence of discrimination across time: demographic variables significantly impact pricing to this day, with birthplace remaining the main discriminatory factor against individuals not born in Italian cities. We also found that driver profiles can determine the number of quotes available to the user, denying equal opportunities to all.

Conclusion. The study underscores the importance of testing for non-discrimination in software systems that affect people's everyday lives. Performing algorithmic audits over time makes it possible to evaluate the evolution of such algorithms. It also demonstrates the role that empirical software engineering can play in making software systems more accountable.

Keywords: algorithmic bias, fairness, software audit, empirical methods

1 Introduction and Motivation

As the level of integration of software systems into the infrastructure of society rapidly increases, so does the responsibility of software professionals [1]. They must ensure that the software they deploy meets a wide range of non-functional requirements such as security, safety, privacy, and, increasingly, non-discrimination. Addressing issues of unfairness in different application domains has become a critical and urgent task: recent studies have identified algorithmic discriminations in areas that are crucial for maintaining equal opportunities for all in society, such as education [2], healthcare [22], employment [19], justice [21] and commerce [12]. In the context of the software engineering discipline, testing software for non-discrimination is a form of non-functional testing [11,3], since biased and unfair decisions have a serious impact on software quality of use, where societal and ethical risks have been recognised in recent standards developments [16,17] and scientific proposals [29]. In addition, the introduction of the AI Act includes proof of non-discrimination for high-risk systems [6].

This study is an audit of an online software system in the Italian car insurance industry: it is ethically and legally imperative to ensure that the software automates and differentiates pricing equitably. We update and extend an original paper by Fabris et al. [9], which quantified the impact of gender and birthplace on insurance pricing in an online system. We extend the original study by including additional demographic attributes, which results in a larger and more nuanced test protocol. This research highlights the importance of testing for non-discrimination in automated pricing algorithms and, more broadly, in software systems that — as the ACM Code of Ethics and Professional Conduct recognises — are "*integrated with everyday activities such as commerce, travel, government, healthcare and education*" [1] (Article 3.7).

The remainder of the paper is organized as follows: in Section 2, we frame the regulations for the domain of interest, and we describe related work in the field; in Section 3, we describe the research questions and experiment design; we report on results in Section 4 and discuss them in Section 5; in Section 6, we discuss the potential threats to validity; finally, in Section 7, we summarize the findings and explore future research directions.

2 Background and related work

The Italian car insurance market is regulated by the Italian Insurance Supervisory Authority (IVASS), which collaborates with international organizations like the International Association of Insurance Supervisors (IAIS) and the European Insurance and Occupational Pensions Authority (EIOPA) to ensure market stability and compliance with international standards. This framework prioritizes consumer protection and prohibits discriminatory practices in premium calculations based on personal characteristics. As a matter of fact, following a 2011 ruling by the European Court of Justice [7], insurers cannot use gender when setting rates. The Article 16 of the European Directive 2021/2118 [5] stipulates that no premium surcharges may be applied on the basis of nationality.

The role of comparison websites is to act as an intermediary between customers and insurance providers. Generally, the insurance companies cover the charges while the customers use the services for no fee. Comparison websites are crucial in the Italian insurance market: during 2022, 54% of people who took out car or motorbike insurance consulted a comparator[1]. They mediate access to insurance not only by aggregating insurance options, but also by customizing offers by algorithmic optimization. Ongoing monitoring of these websites is

[1] https://assicurazioni.segugio.it/news-assicurazioni/polizze-auto-e-moto-la-comparazione-si-conferma-lo-strumento-di-riferimento-degli-italiani-00037593.html

essential to ensure they provide unbiased and accurate information, fostering a fair and competitive market.

Algorithmic audits draw from social sciences and can be defined as the collection and analysis of outcomes from an algorithm and a system, to evaluate its accountability [14]. The audit typically simulates a mock user population, with the objective of finding undesired patterns in models of interest [28]. Algorithmic audits have been frequently used to identify potential discrimination by AI systems. A largely influential work in the field of algorithmic audits is that conducted by Raji and Boulamwini, who described Gender Shades, an algorithmic audit of gender and skin type performance disparities in commercial facial analysis models [23]. The study was replicated after three years by the same authors [24], highlighting the importance of revisiting the algorithms even long after the original audits have been performed. Other domains have been interested in algorithmic audits: e.g., Liu et al. have applied auditing to AI applications in the medical domain, to uncover potential algorithmic errors in the context of a clinical task and anticipate their potential consequences [20]. The auditing process can also involve other techniques than the simulation of inputs. Shen et al. have described the concept of everyday user algorithmic auditing, a process in which users detect, understand, and interrogate problematic machine behaviours via their day-to-day interactions with algorithmic systems [26]. Various institutions have introduced bias auditing systems, as in the case of the New York City Council [27], reinforcing the need for clearer definitions and metrics [15]. In general, the importance of the audit process in exposing the limitations of deployed systems is widely recognised [4].

The foundation of this research stems from a 2021 study [9], which delves into the Italian car insurance industry. This work uncovers significant insights into the influence of *gender* and *birthplace* on car insurance pricing. Despite regulations prohibiting discriminatory practices, the audit revealed that algorithmic-mediated prices are still influenced by these factors. Specifically, foreign-born drivers and individuals from certain Italian cities faced price disadvantages, with Laos drivers being charged up to 1000 € more than drivers with similar profiles in Milan. Additionally, user profiles labelled by the platform as risky received fewer quotes. Building upon this foundational work, the present article describes the evolution of discriminatory and opaque practices in car insurance pricing.

3 Methodology

This section outlines the research methodology used in this study: firstly we illustrate the research questions, then we report on the experiment design. Finally, we describe the data collection process.

3.1 Goal and research questions

The goal of the research is defined with the GQM (Goal-Question-Metric [25]) template as follows: **test** pricing algorithms **for the purpose of** auditing an online software system **with respect to** non-discrimination **from the point of view** of the users **in the context of** the Italian car insurance industry. Two research questions stem from this goal.

RQ1: Do protected attributes (*gender*, *birthplace*, *age*) and socio-demographic attributes (*city*, *marital status*, *education*, *profession*) directly influence quoted premiums? This question would determine whether protected attributes[2] and socio-demographic attributes are related to the insurance quotes provided to the users. This question aims at investigating whether prices vary for similar profiles that differ in only one attribute.

RQ2: Do protected attributes (*gender*, *birthplace*, *age*), socio-demographic attributes (*city*, *marital status*, *education*, *profession*) and driving attributes (*car*, *km driven*, *class*) influence the number of quotes presented to the user? This question focuses on how often car insurance companies appear in search results for different driver profiles. The rationale is to investigate whether profile characteristics expose users to fewer offers, suggesting discrimination when it comes to the availability of quotes among insurance options for them.

3.2 Analysis method

We performed a preliminary exploratory analysis by analysing the prices' distribution for each attribute value. As an example, we assessed the spread between the average premium for male and female drivers. The preliminary exploratory analysis is presented in Section A of the Supplementary materials[3].

Discrimination Analysis (RQ1) To investigate whether protected attributes directly influence quoted premiums, we analysed the distribution of the price differences δ for pairs of profiles that differ only in one attribute. The statistical reliability of the median was checked using a sign test with $\mu_0 = 0$ and considering significant p-values below the 0.05 threshold. We analysed the top1 and the top5 quotes. The top1 analysis concentrates exclusively on the most affordable quote for each profile: this is particularly relevant from the viewpoint of a person who is primarily concerned with the lowest insurance cost. In contrast, the top5 analysis compares the averages of the five cheapest quotes for every profile, catching a deeper view of the insurance rates for a specific profile.

Output Variability (RQ2) To investigate whether specific attributes directly influence the number of quoted premiums offered to users, we observed the proportion of profiles for which each company was present in the offers provided by the comparison site. E.g., to examine the presence of a company for *gender*=male, we calculated the ratio of male profiles for which the company offered a quote to the total number of male profiles. By computing this ratio, we verify whether an attribute influences the number of quotes presented to the users.

[2] Protected attributes are selected among the ones identified through Article 21 *"Non-discrimination"* of the EU Charter of Fundamental Rights [8].

[3] https://anonymous.4open.science/r/RCA-audit-D049/supplementary_materials.pdf

Table 1. Features under analysis. Italics highlight the difference from the original study. Abbreviations are shown in brackets. Differences are motivated in Section 3.3.

Feature	**Values tested in the original study**	**Values tested in this study**
Gender	Male, Female	Male, Female
Birthplace	Milan, Rome, Naples, *Romania*, *Ghana*, *Laos*	Milan (MI), Rome (RO), Naples (NA), *Morocco (MA)*, *China (CN)*
Age	*18*, 25, 32	25, 32
City	Milan, *Rome*, Naples	Milan (MI), Naples (NA)
Marital Status	*Not present*	*Married (Mar), Single (Sin), Widow (Wid)*
Educational qualification	*Not present*	*Master (MSc), Without a qualification (Waq)*
Profession	*Not present*	*Employee (Emp), Looking for a job (Lfaj)*
Car type	Old, Large Engine, Diesel; New, Small Engine, Petrol	Old, Large Engine, Diesel (OLED); New, Small Engine, Petrol (NSEP)
Km traveled in one year	10000, 30000	10000, 30000
Class	1, 4, 9, *14*, 18, *None*	1, 4, 9, 18
Total queries	2160	7680

Table 2. Number of quotes (#Q) and frequency of quotes (Freq.) for each company (Comp.) in the collected data.

Comp.	**#Q**	**Freq.**	**Comp.**	**#Q**	**Freq.**	**Comp.**	**#Q**	**Freq.**	**Comp.**	**#Q**	**Freq.**
C1/a	7.680	100%	C2/a	2.818	36%	C3/b	482	6%	C5/a	5.853	76%
C1/b	1.831	24%	C2/b	1.960	24%	C3/c	1.844	24%	C5/b	1.825	24%
C1/c	7.660	99%	C2/c	177	3%	C3/d	1.419	19%	C6/a	499	6%
C1/d	2.792	35%	C3/a	1.894	25%	C4/a	3.912	64%			

3.3 Experiment Design and Data Collection

The data source is a popular italian comparison website. We collected about 700 insurance policy quotes per day for two weeks, in the second half of January 2024, for a total of 7680 queries. Automatic web scraping with IP rotation with different proxies was integrated by manual data collection. Three potential sources of variability could be identified: the evolution of insurance company models and pricing schemes, the session duration effects, and the A/B testing conducted by insurance companies and/or by the online software itself. To ensure the robustness of the data collected, we used a double nested randomization approach [9]. The first level of randomization (inner loop) involved randomizing the order in which profiles were queried on the website, so that each profile combination (e.g. birthplace RO and gender M) had an equal chance of being queried at any given time. The second level of randomization (outer loop) involved randomizing the order of query blocks, as each block consisted of a set of profiles that were to be queried together.

A control group was included in the study design. We collected control pairs performing two identical queries. This group was expected to not be affected by the variables being tested, and so can be used as a baseline against which the results of the profiles could be compared. If the price differences for protected pairs are significantly different from those of control pairs in terms of *frequency* (percentage of records with a difference less than ±5 €, i.e. Ties_5) and *magnitude* (assessable by observing percentiles), this indicates a bias in the insurance pricing mechanism. Control pairs are also helpful to manage the risk of having price fluctuations related to non-modelled factors (e.g. price updates).

We report a summary of the variables used in Table 1, comparing them with the original study. We removed some values analysed in the original paper because, despite IP address rotation, the platform blocked automatic data collection and web tests were integrated with manual tests. We have removed the less relevant values: *Rome* as place of residence; *age* 18, because at that age in Italy you can only drive a car with a low power to weight ratio; *Romania* because it is an EU country; *class* 14 because in the original study it turned out to represent an intermediate value between classes 9 and 18. We have changed the two non-EU countries to MA and CN because people with these nationalities make up the largest African and Asian (respectively) communities in Italy[4].

We have released the data collected and the Python scripts used to analyse them in an open-source repository [5].

4 Results

The data collected from the queries include six different companies that were offering up to four insurance products for each query. The insurance companies are labelled progressively from C1 to C6, with the corresponding services being differentiated as /a, /b, /c, /d. Different services represent different commercial offers from the same company (e.g. with or without GPS tracking). Examination of the compiled insurance quotes demonstrates the divergence of frequencies and number of quotes offered by each insurance company and product set (details reported in Table 2).

4.1 Discrimination Analysis (RQ1)

Tables 3 and 4 presents the results of the discrimination analysis. Negative values indicate that profiles with the test value obtained lower prices; positive values indicate that they obtained higher quotes compared to those obtained by the baseline. We observe that:

- *Gender*, confronting females and males profiles, on one hand shows a median equal to zero in both top1 and top5. On the other hand, the difference

[4] istat.it - Italy, regions, provinces - Country of citizenship (Frequency: Annual, Indicator: Foreign census population on 1st January, Time: 2023)

[5] https://anonymous.4open.science/r/RCA-audit-D049/

Table 3. Discrimination analysis Top1. The first value of each pair is the *test value*, while the second value of each pair is the *baseline*. The difference prices are calculated as test minus baseline. Ties_5 represents the percentage of protected pairs for which quote difference is within a tolerance threshold of ±5 euros; $\eta.05(\delta)$ and $\eta.95(\delta)$ the 5th and the 95thy percentiles, respectively; $\eta.50(\delta)$ is the median; $m(\delta)$ represents the mean; the p-value tests the null hypothesis that the median difference is zero.

		Top1					
Attribute	Pairs	Ties_5	$\eta.05(\delta)$	$\eta.50(\delta)$	$\eta.95(\delta)$	$m(\delta)$	p
Gender	F vs M	89%	-14 €	0 €	6 €	4 €	0.72
Birthplace	RO vs MI	25%	0 €	8 €	10 €	7 €	<**0.05**
Birthplace	NA vs MI	10%	-8 €	29 €	538 €	92 €	<**0.05**
Birthplace	MA vs MI	0%	-255 €	125 €	539 €	148 €	<**0.05**
Birthplace	CN vs MI	10%	-104 €	103 €	395 €	118 €	<**0.05**
Age	25 vs 32	1%	-1 €	64 €	549 €	141 €	<**0.05**
City	NA vs MI	31%	-255 €	147 €	1752 €	367 €	<**0.05**
Mar. Stat.	Sin vs Mar	79%	-33 €	0 €	83 €	-1 €	<**0.05**
Mar. Stat.	Wid vs Mar	79%	-30 €	0 €	183 €	9 €	<**0.05**
Education	WaQ vs MSc	77%	-9 €	0 €	491 €	60 €	<**0.05**
Profession	LfaJ vs Emp	70%	-377 €	0 €	283 €	24 €	<**0.05**
Control pairs (noise)		98%	0 €	0 €	0 €	7 €	1.00

between $\eta.95(\delta)$ and $\eta.05(\delta)$ raises from 20 € in top1 to 189 € in top5 (0 € in control pairs), while $Ties_5$ drop from 89% to 78% (98% in control pairs).

- *Birthplace*, is used to the advantage of drivers born in Milan. Drivers born outside Italy were the most highly rated: on average, drivers born in Morocco are charged 200 € more than otherwise identical drivers born in Milan.
- *Age* shows a median difference of 64 € in the analysis of the cheapest offer, this value increases to 211 € when looking at the average of the five cheaper offers.
- The *city* attribute reveals better prices in favour of Milan residents. The median and mean values show greater discrimination between Naples and Milan for the City attribute than for the *birthplace* attribute.
- *Marital status*: The top1 values show a median and an average equal or close to zero for both pairs, while the 95th percentile shows values 3 to 6 times higher than the 5th percentile. In the top5 analysis, the median and the average are higher than zero, especially for the pair 'widowed vs. married' (35 € and 110 € respectively).
- *Education* shows a relevant imbalance in the top5 analysis for the damage to the unqualified (WaQ), with a price difference that is 99 € higher on the median and 236 € higher on the average. For the top1 values this effect is reduced: the median is 0 €, while the average is 60 €.
- For *Profession*, analysing the top5 values, jobseekers (Lfaj) received offers that were on median 22 € higher and on average 135 € higher.

Taking into account the significant disparities directly caused by protected attributes, we observe that the magnitude of differences in the top5 results,

Table 4. Discrimination analysis on the averaged five cheapest values (columns descriptions in Table 3).

		Top5					
Attribute	Pairs	Ties_5	$\eta.05(\delta)$	$\eta.50(\delta)$	$\eta.95(\delta)$	$m(\delta)$	p
Gender	F vs M	78%	-61 €	0 €	128 €	4 €	<**0.05**
Birthplace	RO vs MI	71%	0 €	2 €	9 €	4 €	<**0.05**
Birthplace	NA vs MI	4%	-199 €	113 €	382 €	128 €	<**0.05**
Birthplace	MA vs MI	0%	-178 €	252 €	1187 €	371 €	<**0.05**
Birthplace	CN vs MI	4%	-75 €	128 €	712 €	200 €	<**0.05**
Age	25 vs 32	1%	0 €	211 €	877 €	285 €	<**0.05**
City	NA vs MI	23%	-346 €	278 €	1954 €	657 €	<**0.05**
Mar. Stat.	Sin vs Mar	26%	-116 €	10 €	238 €	42 €	<**0.05**
Mar. Stat.	Wid vs Mar	25%	-56 €	35 €	515 €	110 €	<**0.05**
Education	WaQ vs MSc	29%	-7 €	99 €	896 €	236 €	<**0.05**
Profession	LfaJ vs Emp	24%	-188 €	22 €	769 €	135 €	<**0.05**
Control pairs (noise)		98%	0 €	0 €	0 €	10 €	1.00

measured as $\eta.95(\delta) - \eta.05(\delta)$, vary from 9€ (birthplace Rome vs Milan) to 2300€ (city Naples vs Milan), compared to a value of 0 € for control pairs. The frequency of Ties5 for top5 is below 5% for age and all the birthplace pairs except for Rome vs Milan, compared against a value of 98% for control pairs. Similar, albeit slightly less pronounced, patterns are observed in the top1 results.

RQ1: Protected attributes (*gender*, *birthplace*, *age*) and socio-demographic attributes (*city*, *marital status*, *education*, *profession*) have a direct effect on premiums quoted. *Birthplace*, *age*, *city* and *education* systematically reveal a direction of such influence.

4.2 Output Variability (RQ2)

Figure 1 provides an overview of the percentage of quotes presented by each company, for each value of the attributes. Companies *C1* and *C5* are not represented, as they submitted at least one offer for all profiles, i.e. they reached 100% for each class of all attributes. Regarding the other companies: *C2* appeared for about a third of the profiles, *C3* for about a quarter of the profiles, *C6* offered an insurance only for a small subset of records and *C4* was present in about 60% of the queries. It is possible to see a few relevant differences for different profiles, but these differences are company-specific with no general trends:

- The company *C2* submitted fewer quotes for profiles with *birthplace* setted to Morocco. At the same time, it submitted quotes for about 75% of profiles with the car type Old, Large Engine, Diesel (OLED), while no quotes appeared for New, Small Engine, Petrol (NSEP).

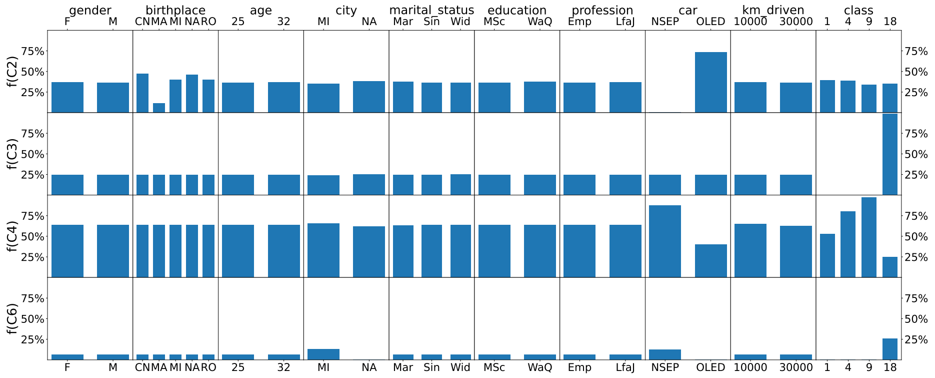

Fig. 1. Presence of companies as a percentage of offers for each attribute.

- The company *C3* was rather balanced in all characteristics, except for risk *class*: it was present in almost all profiles with the highest risk class (18), but was absent for the others.
- The company *C4* presents significantly more quotes for NSEP cars than for OLED cars. It also presents a non-uniform distribution of quotes per risk class.
- The company *C6* presents fewer quotes, so it is more difficult to draw relevant conclusions. However, it is possible to notice that it only presented quotes for the riskier driving class and for profiles living in Milan, and that there are no quotes for OLED cars.

RQ2: *Birthplace* and *city* are factors that influence the number of quotes in two companies. Driver characteristics such as *car* type and *risk class* also expose users to a different number of quotes in four companies.

5 Discussion

RQ1: Do protected attributes (*gender*, *birthplace*, *age*) and socio-demographic attributes (*city*, *marital status*, *education*, *profession*) directly influence quoted premiums? The original study looked at price discrimination based on *birthplace* and *gender*. With regard to the first one, we confirm that the *birthplace* has a direct impact on the premium offered: people born in Naples systematically receive higher prices. This type of discrimination is even more pronounced when we look at the profiles with an Italian birthplace compared to the profiles with a foreign birthplace (in our case, Moroccan and Chinese). With regard to *gender*, as in the original study, there is no systematic advantage in one direction (male or female). *Age* is used to systematically discriminate against younger people: this is expected as – on average – the youngest have a higher risk class, however, this is not systematic as the risk class in Italy can be inherited from a family member. Further investigations looking at the intersection of age and risk class are needed. *City* of residence is the attribute that reveals the greater systematic discrimination: residents in Milan pay, on average, 128 € less than residents in Naples. On the one hand, it seems that insurance companies consider driving in Naples to be riskier than driving in Milan. On the other hand, the Italian National Statistics Institute, in its latest report on road accidents (2022), showed that the absolute number of road accidents and the fatality rate per 100,000 inhabitants were higher in Milan than in Naples [18]. *Marital status* and *profession* show a slight discrimination against married, widowed and unemployed profiles. In terms of *education*, those with no qualifications are paying way much more than those with a Master's degree (in the top5 analysis). We can hypothesise that these attributes play the role of a proxy variable for the age attribute, since younger people are considered to be more at risk of car accidents [13]. However, further multivariate distribution analysis should be performed to confirm this hypothesis. Nevertheless, the magnitude of the differences confirms that all the protected characteristics analysed play a relevant role in the pricing algorithms.

RQ2: Do protected attributes (*gender*, *birthplace*, *age*), socio-demographic attributes (*city*, *marital status*, *education*, *profession*) and driving attributes (*car*, *km driven*, *class*) influence the number of quotes presented to the user? The original study showed that younger people, residents of Naples and drivers with the worst risk class received fewer offers. With regard to *age*, we do not observe any significant differences between 25 and 32, but it should be noted that we did not test the age of 18 (as described in Section 3.3), which seemed to be the least desirable for insurance companies. Observing the influence of the city of residence, only one company confirms the pattern of imbalance between Milan and Naples in terms of the number of offers. As for the claims' history (*class*), the results are controversial. On the one hand, company *C4* confirms the pattern shown in the original study, i.e. the fact that the riskier profiles receive fewer quotes. On the other hand, companies *C3* and *C6* appear only for the highest risk class. We can hypothesise that some companies are attracted by the possibility of obtaining higher revenues.

6 Threats to Validity

In this section, we report the main threats to the validity of the study according to classifications available in the literature [10]. The main *external validity threat* of our study is that the examination has been conducted on only one comparator website. Future studies shall expand the scope to include multiple systems to provide a more comprehensive understanding of the issues. It is not ensured as well whether our results can be applied to other countries or other domains. As *internal validity threats* are concerned, the major ones lie in the selection of the variables that were analysed. Additional variables among those collected by the algorithms could be in fact investigated, including further proxies of socio-economic conditions and intersectional attributes.

7 Conclusions and future work

This study is an extended and updated audit of pricing algorithms used in an online system comparing car insurance prices in Italy. The analysis confirms and extends the findings of the original study: several demographic variables had a significant impact on pricing, with place of birth emerging as a discriminatory factor, especially for those not born in Italian cities. In terms of non-equal opportunities, driver profiles (e.g., car type) determine the options available to users, and for two companies this was based on protected attributes.

These results not only demonstrate the importance of verifying fairness in algorithmic pricing mechanisms and, more generally, of continuously testing software services for non-discrimination, but they also show that empirical methods – especially in relation to experiment design in conjunction with testing techniques – are well suited to such goals. As a result, our work is an example of the potential role that empirical software engineering can play in the emerging field of testing algorithms for non-discrimination. With this vision in mind, future work in this domain will be devoted to increasing the level of test automation, switching features – including intersectional attributes– and tracking prices over time. More generally, we hope that this vision lead to future experiments in the other areas identified by Art. 3.7 of the ACM Code of Ethics and Professional Conduct, where non-discrimination is paramount.

References

1. ACM Code 2018 Task Force: ACM Code of Ethics and Professional Conduct (2018), https://www.acm.org/code-of-ethics
2. Baker, R.S., Hawn, A.: Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education **32**(4), 1052–1092 (Dec 2022). https://doi.org/10.1007/s40593-021-00285-9
3. Brun, Y., Meliou, A.: Software fairness. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 754–759. ESEC/FSE 2018, Association for Computing Machinery, New York, NY, USA (Oct 2018). https://doi.org/10.1145/3236024.3264838
4. Conitzer, V., Hadfield, G.K., Vallor, S.: Technical Perspective: The Impact of Auditing for Algorithmic Bias. Communications of the ACM **66**(1), 100 (Dec 2022). https://doi.org/10.1145/3571152
5. Council of the European Union: Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Jun 2000), http://data.europa.eu/eli/dir/2000/43/oj/eng
6. Deck, L., Müller, J.L., Braun, C., Zipperling, D., Kühl, N.: Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness (Mar 2024). https://doi.org/10.48550/arXiv.2403.20089
7. European Court of Justice: Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres (Mar 2011), https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62009CJ0236

8. European Union Agency For Fundamental Rights: EU Charter of Fundamental Rights - Title III: Quality - Article 21 - Non-discrimination (Apr 2015), http://fra.europa.eu/en/eu-charter/article/21-non-discrimination
9. Fabris, A., Mishler, A., Gottardi, S., Carletti, M., Daicampi, M., Susto, G.A., Silvello, G.: Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 458–468. AIES '21, Association for Computing Machinery, New York, NY, USA (Jul 2021). https://doi.org/10.1145/3461702.3462569
10. Feldt, R., Magazinius, A.: Validity threats in empirical software engineering research-an initial survey. In: Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering. pp. 374–379. Knowledge Systems Institute Graduate School, Redwood City, San Francisco Bay, CA, USA (2010), https://www.cse.chalmers.se/~feldt/publications/feldt_2010_validity_threats_in_ese_initial_survey.pdf
11. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: Testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. pp. 498–510. ESEC/FSE 2017, Association for Computing Machinery, New York, NY, USA (Aug 2017). https://doi.org/10.1145/3106237.3106277
12. Gautier, A., Ittoo, A., Van Cleynenbreugel, P.: AI algorithms, price discrimination and collusion: A technological, economic and legal perspective. European Journal of Law and Economics **50**(3), 405–435 (Dec 2020). https://doi.org/10.1007/s10657-020-09662-6
13. Gomes-Franco, K., Rivera-Izquierdo, M., Martín-delosReyes, L.M., Jiménez-Mejías, E., Martínez-Ruiz, V.: Explaining the Association between Driver's Age and the Risk of Causing a Road Crash through Mediation Analysis. International Journal of Environmental Research and Public Health **17**(23), 9041 (Jan 2020). https://doi.org/10.3390/ijerph17239041
14. Goodman, E.P., Trehu, J.: ALGORITHMIC AUDITING: CHASING AI ACCOUNTABILITY. Santa Clara High Technology Law Journal **39**(3), 289 (May 2023), https://digitalcommons.law.scu.edu/chtlj/vol39/iss3/1
15. Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., Strait, A.: Auditing Work: Exploring the New York City algorithmic bias audit regime (Feb 2024). https://doi.org/10.48550/arXiv.2402.08101
16. ISO: ISO/IEC 25019:2023 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality-in-use model. Standard, International Organization for Standardization, Geneva, CH (2023), https://www.iso.org/standard/78177.html
17. ISO: ISO/IEC 25059:2023 - Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems. Standard, International Organization for Standardization, Geneva, CH (2023), https://www.iso.org/standard/80655.html
18. Istituto Nazionale di Statistica: Incidenti stradali in Italia. Anno 2022. Tech. rep., ISTAT (2022), https://www.istat.it/it/archivio/286933
19. Köchling, A., Wehner, M.C.: Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research **13**(3), 795–848 (Nov 2020). https://doi.org/10.1007/s40685-020-00134-w
20. Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K., Oakden-Rayner, L.: The medical algorithmic audit. The Lancet Digital Health **4**(5), e384–e397 (May 2022). https://doi.org/10.1016/S2589-7500(22)00003-6

21. Malek, M.A.: Criminal courts' artificial intelligence: The way it reinforces bias and discrimination. AI and Ethics **2**(1), 233–245 (Feb 2022). https://doi.org/10.1007/s43681-022-00137-9
22. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (Oct 2019). https://doi.org/10.1126/science.aax2342
23. Raji, I.D., Buolamwini, J.: Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 429–435. AIES '19, Association for Computing Machinery, New York, NY, USA (Jan 2019). https://doi.org/10.1145/3306618.3314244
24. Raji, I.D., Buolamwini, J.: Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Communications of the ACM **66**(1), 101–108 (Dec 2022). https://doi.org/10.1145/3571151
25. Rini van Solingen, Basili, V., Caldiera, G., Rombach, H.D.: Goal Question Metric (GQM) Approach. In: J.J. Marciniak (ed.) Encyclopedia of Software Engineering. John Wiley & Sons, Ltd, USA (2002). https://doi.org/10.1002/0471028959.sof142
26. Shen, H., DeVos, A., Eslami, M., Holstein, K.: Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 433:1–433:29 (Oct 2021). https://doi.org/10.1145/3479577
27. The New York City Council: A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools (2021), https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=
28. Vecchione, B., Levy, K., Barocas, S.: Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–9. EAAMO '21, Association for Computing Machinery, New York, NY, USA (Nov 2021). https://doi.org/10.1145/3465416.3483294
29. Vetrò, A., Torchiano, M., Mecati, M.: A data quality approach to the identification of discrimination risk in automated decision making systems. Government Information Quarterly **38**(4), 101619 (Oct 2021). https://doi.org/10.1016/j.giq.2021.101619