

# Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness

Luca Deck<sup>1,2</sup>, Jan-Laurin Müller<sup>1</sup>, Conradin Braun<sup>1</sup>, Dominique Zipperling<sup>1,2</sup> and Niklas Kühl<sup>1,2</sup>

<sup>1</sup>University of Bayreuth

<sup>2</sup>Fraunhofer FIT

## Abstract

The topic of fairness in AI, as debated in the FATE (Fairness, Accountability, Transparency, and Ethics in AI) communities, has sparked meaningful discussions in the past years. However, from a legal perspective, particularly from the perspective of European Union law, many open questions remain. Whereas algorithmic fairness aims to mitigate structural inequalities at design-level, European non-discrimination law is tailored to individual cases of discrimination after an AI model has been deployed. The AI Act might present a tremendous step towards bridging these two approaches by shifting non-discrimination responsibilities into the design stage of AI models. Based on an integrative reading of the AI Act, we comment on legal as well as technical enforcement problems and propose practical implications on bias detection and bias correction in order to specify and comply with specific technical requirements.

## Keywords

EU AI Act, Algorithmic fairness, Non-discrimination law, Ethical AI

## 1. Introduction

AI systems' propensity to discriminate against legally protected groups has been demonstrated across multiple social contexts, ranging from decision-support systems for criminal risk assessment [1], recruiting [2], and credit scoring [3], to applications in computer vision [4, 5, 6, 7] and natural language processing [8, 9, 10]. In light of the rapid advancements of AI, the increasing use of AI systems across multiple domains has triggered a broad and interdisciplinary debate on the "ethics of algorithms" [11, 12, 13]. Central to this debate are the FATE principles (fairness, accountability, transparency, and ethics), with fairness encompassing the social goals of non-discrimination, inclusion, and equality [14, 15].

The discourse at the interface with legal scholarship, however, is only starting to gain traction (e.g., [16, 17, 18, 19, 20, 21]). In this short paper, we make three contributions: First, we briefly retrace the academic discourses on non-discrimination law and algorithmic fairness to highlight their current misalignment. Second, we argue that the European Union's AI Act might pose a seminal link to merging these debates. Based on this integrative conception, we thirdly


---

*EWAf'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany*

✉ luca.deck@uni-bayreuth.de (L. Deck); jan-laurin.mueller@uni-bayreuth.de (J. Müller); conradin.braun@uni-bayreuth.de (C. Braun); dominique.zipperling@uni-bayreuth.de (D. Zipperling); kuehl@uni-bayreuth.de (N. Kühl)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

sketch how the AI Act could provide a means to solve the enforcement problems of both—non-discrimination law and algorithmic fairness—and comment on upcoming challenges regulators and developers will face when specifying and verifying technical requirements.

## 2. Non-discrimination law vs. algorithmic fairness

**Legal Context: Non-discrimination law and its shortcomings.** From a legal perspective, non-discrimination law appears to be suitable to address the potential harms of unfair AI systems at first glance. However, legal scholars from both sides of the Atlantic have demonstrated that U.S. [22] and EU [16] non-discrimination law alike may fall short in doing so. One of the main deficiencies of traditional non-discrimination regimes in the context of algorithmic discrimination is law enforcement. Enforcement has always been a central shortcoming of non-discrimination law, especially in jurisdictions that primarily rely on individual litigation (cf., [23, 24]). In such jurisdictions, individual victims face substantial problems when it comes to recognizing, proving, and bringing instances of discrimination before the courts. AI systems exacerbate these problems [25]. Due to the opacity of these systems, those affected by algorithmic discrimination are often unable to recognize instances of (potential) discrimination [26]. Moreover, even when individuals suspect discrimination, restricted access to models or training data severely impedes their ability to meet the requirements of the burden of proof imposed on them by procedural law [16]. Furthermore, European non-discrimination law is tailored to individual cases of discrimination hampering its application to broad-scale goals like designing fair AI systems. Non-discrimination regimes, therefore, face substantial challenges when it comes to enforcing the principles of equality and non-discrimination.

**Technical context: algorithmic fairness and its shortcomings.** On a technical level, methods for algorithmic fairness from the field of computer science set out to fill this gap. By developing a plethora of technical bias definitions and fairness metrics (cf. [27, 28, 29]) as well as practical bias detection and bias mitigation techniques [30, 31, 32, 33], computer scientists try to implement ethical and legal fairness considerations “by design” [34]. The shortcomings of these technical fairness approaches, however, are twofold: First, formalization and quantification will never provide answers to fundamentally normative challenges such as selecting the right fairness metric for the right context or trading off conflicting objectives [35, 36]. Such challenges arising from conflict between values can be supported but not be solved by formal methods [37]. Second, due to its orientation towards a specific academic audience and reliance on self-governance, discourse on algorithmic fairness faces its own “enforcement problems” [38]. The AI Act may alleviate both—the enforcement problems of non-discrimination law and the technical fairness discourse—alike.

## 3. Implications of the AI Act

**Enforcement “by design”?** According to Recital 4a, the AI Act explicitly aims to protect the fundamental rights set out in Art. 2 of the Treaty of the European Union. Among these rights are equality and non-discrimination in particular. In order to prevent algorithmic discrimination,

the regulation establishes special requirements (Art. 6 et seq. AI Act) for high-risk systems in the areas of education (Recital 35), employment (Recital 36), insurance and credit (Recital 37), law enforcement (Recital 38), as well as migration (Recital 39). However, despite its explicit goal to prevent discrimination, the regulation lacks a clear substantive standard for determining when unequal treatment is inadmissible. According to Art. 10(2)f AI Act “[t]raining, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the AI system” and thus have to be examined for “possible biases that are likely to [...] lead to discrimination prohibited under Union law”. The AI Act therefore leaves the judgment call about what constitutes illegal discrimination to existing legislation. However, traditional non-discrimination law’s requirements can only be implemented during model development (as intended by the AI Act) if they are “translated” into technical fairness requirements. To achieve this goal, scholars from all domains are bound to collaborate. When doing so, they must proceed in a conscious and contextualizing manner and take into account the diverging perspectives of AI Act and non-discrimination law. European non-discrimination law is tailored to individual instances of discrimination after an AI model has been deployed—an inherently *retrospective* approach. In contrast to this, the AI Act *prospectively* demands fairness interventions by implementing non-discrimination requirements at the stage of model design. Guidance by democratically justified institutions on how to implement such requirements might bridge the gap toward alleviating both the legal and the technical enforcement problems.

**Enabling “bias detection and correction”?** Legal requirements for the development of AI systems are not only subject to the AI Act. Due to the tension between fairness and privacy during the training and evaluation stage of AI, conflicts with data protection law may equally arise. On the one hand, ignoring personal demographic data promotes the same risk as the widely rejected idea of fairness through unawareness because legally protected attributes like race and gender usually correlate to innocuous proxy variables [39, 40]. If protected attributes are unavailable during model training and evaluation, these subtle correlations cannot be accounted for, nor can technical fairness metrics be tested and optimized. On the other hand, Art. 9 GDPR places particularly high demands on the lawful processing of personal data about special categories. Therefore, the same sensitive data that is protected by data protection law is also essential to effectively avoid discriminatory outputs. The AI Act seeks to mitigate this tension by broadening the scope of lawful data processing. Art. 10(5) AI Act states that “[t]o the extent that it is strictly necessary for the purposes of ensuring bias detection and correction in relation to the high-risk AI systems [...], the providers of such systems may exceptionally process special categories of personal data referred to in Art. 9(1) [GDPR].” This is accompanied by Recital 44c, which adds that “[i]n order to protect the right of others from the discrimination that might result from the bias in AI systems [...] the providers should, exceptionally, [...] be able to process also special categories of personal data, as a matter of substantial public interest within the meaning of Art. 9(2)(g) [GDPR].” Therefore, discrimination and fairness considerations can provide a justification for data processing during the training phase of high-risk AI systems. However, balancing the public and private interests regarding non-discrimination and privacy will inevitably lead to intricate trade-offs.

## 4. Practical challenges for compliance

**Defining bias: what are “appropriate” fairness metrics?** The discussed implications of the AI Act raise two important questions on how to put non-discrimination and fairness into practice. First, the concept of technical fairness metrics begs the question which one(s) may be “appropriate for the intended purpose of the AI system” (Art. 10(2)f AI Act). Technical fairness definitions have already been examined for their compatibility with moral norms [41] and non-discrimination regimes [17, 18, 19, 42, 21] alike. However, legal concepts relying on flexible ex-post standards and human intuition are in tension with the mathematical need for precision and ex-ante standardization [21, 42]. Also, the interdisciplinary discourse needs to acknowledge that fairness and non-discrimination might present inherently different concepts targeted at different social contexts. Prior works have suggested that a single standard of fairness can be achieved by “translating” legal non-discrimination requirements from the employment context into technical fairness metrics [17, 19]. However, the heterogeneity of social contexts (e.g., employment versus criminal sentencing) demands a corresponding flexibility in fairness requirements [43, 44]. Instead of aiming for a one-size-fits-all solution, we therefore recommend applying the landscape of available technical fairness metrics to different legal conceptions of discrimination depending on the societal context.

**Detecting and correcting bias: when are biases “likely to lead to discrimination”?** The second challenge is defining when “possible biases that are likely to [...] lead to discrimination”. Technical fairness metrics such as statistical parity or equalized odds offer an actionable approach to measure and mitigate “bias” [45, 30, 21, 46]. However, it remains unanswered what kind of evidence would signal sufficient efforts of bias detection and correction. Setting aside the debate on metric selection, let us assume algorithmic hiring requires male and female applicants to receive equal hiring rates (demographic parity). Statistical hypothesis testing provides a suitable method to verify compliance with this requirement, in this case a simple z-test. To test the hypothesis of compliance with demographic parity, we are interested in the test’s error rates, i.e., falsely detecting a violation (type 1 error) or the likelihood of failing to detect a violation (type 2 error). Notably, a larger disparity in hiring probabilities between groups and a larger sample size decreases type 2 error. Unfortunately, the z-test is also sensitive to the acceptance rate—particularly for small sample sizes. For example, for 1000 male and 1000 female applicants, type 2 error decreases by 0.8% - points if only 700 instead of 900 applicants are accepted—despite identical group disparities (see Appendix A). This effect is especially strong for imbalanced datasets. For 1800 male and 200 female applicants, type 2 error even decreases by 6% - points if only 780 instead of 980 applicants are accepted—again, despite identical group disparities (see Appendix A). Our example highlights the need for guidance in selecting appropriate tests and specifying standards for the error rates of tests utilized in bias detection.

## 5. Conclusion

In this short paper, we outlined how the AI Act could promote the convergence of legal non-discrimination discourse and technical algorithmic fairness discourse. While we sketch its poten-

tial implications on fairness requirements of future AI developments, specifying and enforcing concrete legal requirements will be an intricate future task. In the absence of legal precedents, both disciplines are in need of pioneering work at the intersection of non-discrimination law and algorithmic fairness.

## References

- [1] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks., in: K. Martin (Ed.), Ethics of data and analytics, An Auerbach Book, CRC Press Taylor & Francis Group, Boca Raton and London and New York, 2022, pp. 254–264.
- [2] J. Dastin, Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women \*, in: K. Martin (Ed.), Ethics of data and analytics, An Auerbach Book, CRC Press Taylor & Francis Group, Boca Raton and London and New York, 2022, pp. 296–299.
- [3] A. Fuster, P. Goldsmith-Pinkham, T. Ramadoral, A. Walther, Predictably unequal? the effects of machine learning on credit markets, *The Journal of Finance* 77 (2022) 5–47.
- [4] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, D. Sculley, No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017. URL: <https://arxiv.org/pdf/1711.08536.pdf>.
- [5] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81, PMLR, 2018, pp. 77–91.
- [6] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Men also like shopping: Reducing gender bias amplification using corpus-level constraints, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2979–2989.
- [7] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, A. Rohrbach, Women also snowboard: Overcoming bias in captioning models, 2019. URL: <https://arxiv.org/pdf/1803.09797.pdf>.
- [8] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 4356–4364.
- [9] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* 356 (2017) 183–186.
- [10] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences* 115 (2018) E3635–E3644.
- [11] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, *Big Data & Society* 3 (2016) 1–21.
- [12] M. Kearns, A. Roth, The Ethical Algorithm - The Science of Socially Aware Algorithm Design, Oxford University Press, New York, 2019.
- [13] D. Martens, Data Science Ethics - Concepts, Techniques, and Cautionary Tales, Oxford University Press, New York, 2022.
- [14] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge,

- R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, *Minds and Machines* 28 (2018) 689–707.
- [15] European Commission. Directorate General for Communications Networks, Content and Technology., High Level Expert Group on Artificial Intelligence., Ethics guidelines for trustworthy AI, 2019. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- [16] P. Hacker, Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law, *Common Market Law Review* 55 (2018) 1143–1185.
- [17] S. Wachter, B. Mittelstadt, C. Russel, Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai, *Computer Law & Security Review* 41 (2021) 1–31.
- [18] S. Wachter, B. Mittelstadt, C. Russel, Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law, *West Virginia Law Review* 123 (2021) 735–790.
- [19] M. Hauer, J. H. Kevekordes, M. Amir, Legal perspective on possible fairness measures – a legal discussion using the example of hiring decisions, *Computer Law & Security Review* 42 (2021) 1–20.
- [20] M. Zehlike, P. Hacker, E. Wiedemann, Matching code and law: achieving algorithmic fairness with optimal transport, *Data Mining and Knowledge Discovery* 34 (2020) 163–200.
- [21] H. Weerts, R. Xenidis, F. Tarissan, H. P. Olsen, M. Pechenizkiy, Algorithmic unfairness through the lens of eu non-discrimination law: Or why the law is not a decision tree, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.
- [22] S. Barocas, A. D. Selbst, Big data’s disparate impact, *California Law Review* 104 (2016) 671–732.
- [23] S. Fredman, *Discrimination Law*, 2 ed., Oxford University Press, Oxford, 2011.
- [24] S. Berghahn, V. Egenberger, M. Klapp, A. Klose, D. Liebscher, L. Supik, A. Tischbirek, *Evaluation des Allgemeinen Gleichbehandlungsgesetzes - erstellt im Auftrag der Antidiskriminierungsstelle des Bundes vom Büro für Recht und Wissenschaft GbR mit wissenschaftlicher Begleitung von Prof. Dr. Christiane Brors*, Nomos, Baden-Baden, 2016.
- [25] I. Spiecker gen. Döhmman, E. Towfigh, *Automatisch Benachteiligt. Das Allgemeine Gleichbehandlungsgesetz und der Schutz vor Diskriminierung durch algorithmische Entscheidungssysteme. Rechtsgutachten im Auftrag der Antidiskriminierungsstelle des Bundes*, Antidiskriminierungsstelle des Bundes, Berlin, 2023.
- [26] J. Burrell, How the machine ‘thinks’: Understanding opacity in machine learning algorithms, *Big Data & Society* 3 (2016) 1–11.
- [27] S. Verma, J. Rubin, Fairness definitions explained, in: *Proceedings of the International Workshop on Software Fairness*, ACM Conferences, ACM, 2018, pp. 1–7.
- [28] A. Chouldechova, A. Roth, The Frontiers of Fairness in Machine Learning, *Communications of the ACM* 63 (2018).
- [29] D. Pessach, E. Shmueli, A review on fairness in machine learning, *ACM Comput. Surv.* 55 (2022). URL: <https://doi.org/10.1145/3494672>.

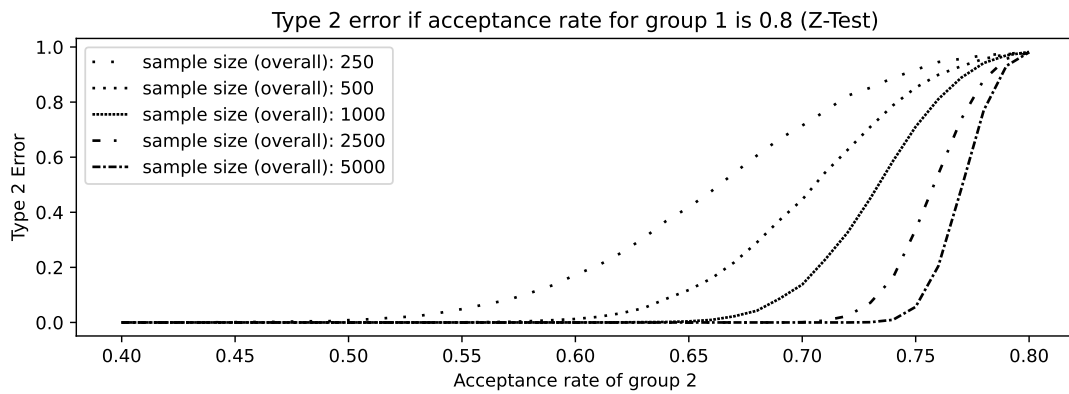


- [30] M. Hardt, X. Chen, X. Cheng, M. Donini, J. Gelman, S. Gollaprolu, J. He, P. Larroy, X. Liu, N. McCarthy, A. Rathi, S. Rees, A. Siva, E. Tsai, K. Vasist, P. Yilmaz, M. B. Zafar, S. Das, K. Haas, T. Hill, K. Kenthapadi, Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021), 2021.
- [31] L. Deck, J. Schoeffler, M. De-Arteaga, N. Kühn, A critical survey on fairness benefits of Explainable AI, in: ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), 2024.
- [32] M. Hort, Z. Chen, J. M. Zhang, M. Harman, F. Sarro, Bias mitigation for machine learning classifiers: A comprehensive survey, *ACM Journal on Responsible Computing* (2023).
- [33] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. S. Oliveira, I. Winkler, E. G. S. Nascimento, Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, *Big Data and Cognitive Computing* 7 (2023) 15.
- [34] I. Žliobaitė, Measuring discrimination in algorithmic decision making, *Data Mining and Knowledge Discovery* 31 (2017) 1060–1089. URL: <https://doi.org/10.1007/s10618-017-0506-1>.
- [35] R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy, *Conference on Fairness, Accountability and Transparency* 81 (2018) 149–159.
- [36] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, The (Im)possibility of fairness, *Communications of the ACM* 64 (2021) 136–143.
- [37] A. Narayanan, The limits of the quantitative approach to discrimination, 2022. URL: <https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/baldwin-discrimination-transcript.pdf>.
- [38] B. Mittelstadt, Principles alone cannot guarantee ethical ai, *Nature Machine Intelligence* 1 (2019) 501–507.
- [39] M. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual Fairness, in: 31st Conference on Neural Information Processing Systems, 2017.
- [40] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12, ACM Press, 2012.
- [41] D. Hellman, Measuring algorithmic fairness, *Virginia Law Review* 106 (2020) 811–866.
- [42] L. Koutsoviti Koumeri, M. Legast, Y. Yousefi, K. Vanhoof, A. Legay, C. Schommer, Compatibility of fairness metrics with eu non-discrimination laws: Demographic parity & conditional demographic disparity, 2023. URL: <https://arxiv.org/pdf/2306.08394.pdf>.
- [43] S. Corbett-Davies, S. Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, 2018. URL: <http://arxiv.org/pdf/1808.00023v2>.
- [44] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, 2020.
- [45] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, [fairmlbook.org](http://fairmlbook.org), 2019.
- [46] Z. Chen, J. M. Zhang, M. Hort, M. Harman, F. Sarro, Fairness testing: A comprehensive survey and analysis of trends, *ACM Transactions on Software Engineering and Methodology*

(2023).

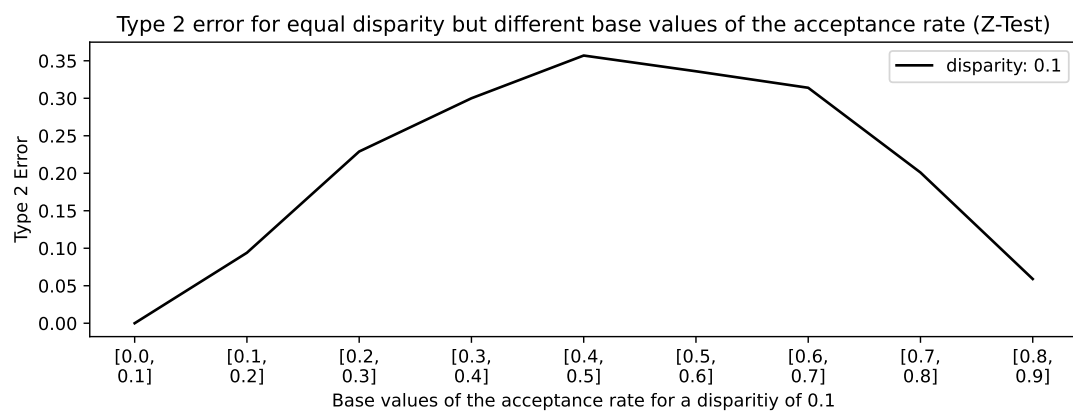
## A. Appendix

The appendix aims to visualize the effects described (Section 4). Figure 1 refers to the effect of larger disparity in hiring probabilities on the probability of not detecting a violation (Type 2 error). For example, for a sample size of 2,500 a change from acceptance rate from 0.75 to 0.7 results in a 17% - point decrease (from 33% to 16%) in type 2 error if group 1 has an acceptance rate of 0.8. Furthermore, it demonstrates that increasing the sample size for the same disparity also decreases the probability of a type 2 error. Doubling the sample size from 2,500 to 5,000 samples decreases the type 2 error by 27% -points (from 33% to 6%). The first effect increases with increasing sample size, while the second one decreases with increasing sample size. Figure 2 demonstrates the effect of the same disparity (0.1) but different acceptance rates. For 1800 male and 200 female applicants, the type 2 error decreases by 6% - points if only 780 (720 male, 60 female) instead of 980 (900 male and 80 female) applicants are accepted. This effect is amplified by imbalanced data sets and small sample sizes.



**Figure 1:** Type 2 error for increasing the disparity of the acceptance rate for two groups





**Figure 2:** Type 2 error for the same disparity in acceptance rate for two groups but for different acceptance rate values