Table of Contents

1. Initial Feature Selection Report

Data & Feature Summary

Our model uses data collected from previous customers, referred to as data generators as well, to calculate the probability of filing a claim and the predicted cost of the claim using classification and regression models. The data contains information about their identity such as age, gender, and marital status, their family information like the number of children they have, their socio-economic data (income, home value, etc.), vehicular information and data on their past claims.

When preprocessing the data, we remove sensitive information (which can lead to biased predictions or privacy violations) about the customers such as the number of children, the value of their home, their income, parental & marital status, gender, education, and occupation. We have only used the top 10 important features in our model, and from this we have excluded the features tagged as sensitive.

Ethical Challenges

In developing our insurance claim prediction models, we face several ethical challenges regarding customer data. Our models require sensitive information including identity details, family information, socioeconomic status, behavioral history, and vehicle information.

From a deontological perspective, we must consider issues of consent, as customers typically provide data for insurance services, not specifically for AI model training. This raises questions about whether blanket consent for "research and product development" is sufficient or if more specific, informed consent is required. There's significant ambiguity about whether individuals maintain rights over their data after sharing it with our company. Frameworks like GDPR [4] suggest individuals retain certain rights over their identity-related information, but the extent of these rights in AI training contexts remains contested. Using sensitive personal information (such as income, family composition, or behavioral history) must be balanced against preserving human dignity, particularly when these attributes influence insurance decisions that could significantly impact customers' financial situations.

From a utilitarianism approach, we recognize that some data could perpetuate historical biases in insurance pricing if used inappropriately. Historical claim data might reflect systemic disparities in claim processing or reporting, potentially leading to discriminatory outcomes when used for prediction. Different data types carry varying levels of sensitivity and potential for harm - vehicle information might be less sensitive than income data, but both influence predictions. The tension between improved insurance modeling benefits (more accurate pricing, better risk assessment, potential cost savings for low-risk customers) and individual privacy rights remains central. While better models could benefit the broader customer base, this must be weighed against privacy concerns.

These ethical challenges require us to develop thoughtful policies that respect individual rights while enabling beneficial innovation. Any approach must balance multiple stakeholders' interests, including customers, regulators, and the company itself while adhering to both legal requirements and ethical principles that may sometimes extend beyond what the law strictly requires.

Proposed Standard

We propose Tiered Consent with Transparency and Purpose Limitation standards for our data collection practices. This approach separates basic insurance service data (required) from research and model development data (opt-in). For model development, we will request separate, explicit consent with plain language explanations, specification of included attributes, and clarification that individual data cannot be fully removed once incorporated into training models.

Our standard emphasizes data minimization by only collecting information with demonstrable relevance to insurance risk. We will establish clear retention periods with subsequent anonymization of personally identifiable information. To maintain transparency, we will issue regular reports showing which data categories influence our predictions.

This standard is ethically permissible because it respects individual autonomy through meaningful consent, minimizes potential harm through purpose limitation, creates transparency around data usage, and balances individual rights with collective benefits while enabling valuable model development. It acknowledges both rights-based and consequence-based ethical frameworks, providing a balanced approach to the complex challenge of using personal data in predictive insurance models.

Potential Objections & Future Solutions

We recognize that using the customer's data for training and building AI modes can raise ethical and legal considerations. While we do not believe that the data generators maintain complete ownership over the results and insights we have derived from the models, we are committed to respecting their privacy and ensuring responsible data usage.

There may be an occurrence where customers may request that their data be removed from the model (post-consent withdrawal). However, we acknowledge that removing individual data points from a large model is practically infeasible. In such cases, we can ensure that the customer data will be stored on the model's server only for the required period of time and will be anonymized or deleted from the server post the period. We can also enable the customer to remove their data from any future use and product development, should they wish for it - under CCPA [1] and CPRA [2], customers can request their data to be removed, but this does not extend to already-trained models. Another concern that the company may face is that the data generators may later claim they misunderstood how their data would be used, leading to disputes. We can address this by using plain, clear language statements and visual aids to explain the data usage. This will allow the customers to opt-in or opt-out of the development processes - under CPRA [3] and GDPR [4], consent must be "freely given, specific, informed and unambiguous".

Customers may object to the collection of their data if they feel that it is irrelevant, such as the number of dependents, their parental status, or the color of their car. In such cases, we can share the categories/features that we will be using in specific models and how that feature may be relevant - under CCPA [5], data collection should be limited to necessary purposes only. This will help to foster trust as well.

There may be a distrust in the data generators due to the opacity of the details of the model, especially when it's regarding the claim costs. We can provide simplified insights to the customers, such as your probability of claim is predicted to be high due to a history of accidents - under CPRA [6], individuals have the right to give meaningful information on automated

decisions. We can also use explainable AI such as LIME to help customers understand the interpretation of the results.

However, we will have to be careful in finding the trade-off line of how much information to divulge to the data generators so that any sensitive data about the company is not at risk. The complete ownership of the results lies with the company and while we try to quench questions about the data usage and practices to the best of our ability, we will ensure that company data will not be at risk of exposure. This approach will address possible objects while ensuring transparent and legally compliant data practices.

2. Explainability Report

Data & Code Overview

In the second update of our model, we have carried forward the insurance data, which includes key demographics, vehicle, and claim history features. We have used the dataset that we had cleaned and filtered in the previous part. In this step, we will be applying a selected explainable AI method to understand and articulate the model's decision logic and the possible counterfactual outcomes for individuals.

Overview of Explainable AI Methods & Ethical Challenges

Explainable AI (XAI) provides critical insights into the complex decision-making processes of machine learning models. These methods are essential for understanding and interpreting algorithmic predictions.

LIME (Local Interpretable Model-agnostic Explanations) generates local explanations for individual predictions by creating a simplified, interpretable model around specific data point. However, it risks oversimplifying complex model decisions and potentially revealing sensitive information about data points or the model's internal workings.

SHAP (SHapley Additive exPlanations) offers a more sophisticated approach with multiple visualization techniques. The bar graph demonstrates overall feature importance, the waterfall plot shows feature contributions with magnitude and direction, and the heatmap visualizes complex feature interactions. Despite its strengths, SHAP faces challenges related to computational complexity and the potential to reinforce existing biases.

Counterfactual XAI methods, such as DiCE (Diverse Counterfactual Explanations), explore alternative scenarios by showing how minimal feature changes could alter model predictions. These methods help users understand model decision boundaries by presenting hypothetical outcomes, though they must carefully avoid suggesting discriminatory modifications.

The ethical landscape of XAI is complex, navigating legal frameworks like GDPR[7] and ECOA[8] while addressing fundamental questions about explanation ownership. Two primary perspectives emerge: a negative rights approach focusing on data privacy, and an inherent rights perspective viewing interpretability as a fundamental aspect of AI systems.

Ethical considerations center on fairness and discrimination. XAI methods must rigorously examine feature importance to detect and prevent potential biases. This involves carefully analyzing how different features contribute to predictions, ensuring the system does not perpetuate existing societal inequalities.

The selection of XAI methods goes beyond technical capabilities. Researchers must consider how these approaches align with ethical standards, preserve individual rights, and provide meaningful insights. The goal is to transform complex algorithmic decisions into clear, understandable explanations that respect privacy and promote transparency.

Ultimately, XAI bridges the gap between advanced machine learning models and human understanding, offering a critical lens through which we can examine and interpret artificial intelligence's decision-making processes.

Explainable AI Methods for Insurance Claim Prediction

In our specific insurance claim prediction context, we strategically selected SHAP (using shap.TreeExplainer) and DiCE as our primary XAI and counterfactual XAI methods. The choice of SHAP is particularly well suited to our tree-based models, providing a robust and theoretically grounded approach to feature importance. Its ability to work effectively with tree-based models aligns perfectly with our insurance prediction framework, offering transparent insights into how different features contribute to claim predictions.

DiCE complements SHAP by generating diverse counterfactual explanations that help stakeholders understand potential scenario variations. In the insurance context, this means providing clear, meaningful insights into how minimal changes in features might affect claim predictions. The method supports our commitment to transparency and ethical data usage, allowing customers to gain insights into the factors influencing their insurance outcomes without compromising individual privacy.

Our selection of these methods reflects a careful balance between technical effectiveness and ethical considerations. By leveraging SHAP and DiCE, we create an approach that is not just analytically powerful but also fundamentally committed to transparency, fairness, and individual rights.

However, while our approach provides clear, instance-level explanations via SHAP and DiCE, which increases user trust and compliance, it comes with its downsides like exposing too much of our model's inner workings. Doing this can make it easier to reverse-engineer the decision boundaries, trick the system into giving favorable decisions or figure out correlations between features. To ensure this does not occur, we will provide only aggregated summaries of the feature importance and enforce strict access controls. This approach will allow us to maintain transparency while protecting our proprietary data methods.

Potential Objections & Future Solutions

While we are incorporating the most suitable xAI methods to offer the best interpretability for the data generators, there may be concerns, both ethical and legal, that they will have. It is good ethical practice for us, as the company, to try to address this in the best legal way possible.

There may be an occurrence where the data generators feel that the xAI method reveals sensitive attributes to them and it would lead to privacy risks. They may raise the concern that removing sensitive features like gender, income, etc might not be sufficient as LIME may highlight correlated features as well. A potential solution that can be applied is to provide the generators' explanations at an aggregate level to reduce the re-identification risk and iterate that data will be collected and modeled following the GDPR [9] and CCPA [10] privacy guidelines.

Another concern to be addressed is that individuals may demand to remove their data and the derived insights from the model and explainability pipeline under the "Right-to-Delete" [11] article from CCPA. We can anonymize the records from the beginning to ensure the protection of

customers' identities and keep training the models regularly to enable us to remove a user's data if requested. We also reiterate and stress that while data can be removed, isolating the learning parameters from an existing model is not feasible.

The local explanation methods can be misleading or oversimplified in the complex models being used and the examples may not always reflect real-life feasibility. We would aim to combine the LIME methodology with global feature importance methods for cross-verification. We clarify that the explanations gained are approximate and will be used as a guiding stick, not a definitive measure, which aligns with the GDPR [12] policy of providing meaningful information.

By anticipating these objections and employing the mitigation strategies, the organization will be able to more effectively balance transparency, user rights, and the protection of proprietary methods.

3. Disparate Treatment Report

Data & Code Overview

In the second update of our model, we have carried forward the insurance data, which includes key demographics, vehicle, and claim history features. In this step, we will understand and articulate the model's decision logic and the possible counterfactual outcomes for individuals. We will define the metrics that we will be excluding from the model to avoid a possible biased/discriminate outcome.

Overview of Schools of Ethics in Discrimination

The ethical frameworks governing discrimination in AI shape how we approach fairness in our insurance model. The Negative Rights/Libertarian approach emphasizes freedom from discriminatory treatment, focusing primarily on removing protected attributes to prevent direct harm. This view aligns with what philosophers like Nozick would consider the "minimal state" of AI ethics — ensuring systems don't actively discriminate [13]. In contrast, the Equal Opportunities/Positive Rights perspective recognizes that simply removing protected features is insufficient when historical inequalities remain embedded in seemingly neutral data. This framework calls for proactive measures to ensure all individuals have equitable outcomes, regardless of background.

Utilitarianism offers a different lens, suggesting we balance overall model performance against potential harms to specific groups. This approach might accept some disparate impact if the aggregate benefit is substantial, though this raises questions about whose utility is prioritized. Finally, the Dehumanization/Dignity perspective (rooted in Kantian ethics) focuses on respecting human autonomy in AI systems through transparency and explainability, rejecting the treatment of people as mere data points [14].

These ethical frameworks inform our standards for feature selection — from simple blindness (removing protected attributes) to more sophisticated approaches like blindness with proxies (removing correlated features), relevance assessment (both statistical and conceptual), and conscientiousness (ongoing monitoring for fairness).

Our Standard

In light of the understanding of the schools of ethics in discrimination, we have decided to exclude a few features from our analysis and model to ensure that there is no bias in the outcomes of the models. We have outlined our rationale for selection of these features based on these principles.

For our model, we initially shortlisted the top 10 most important features of which we decided to drop the features of home value and income. In addition to this, we will also be excluding years on the job (YOJ) in order to prevent historical inequities from influencing predictions. Removing these parameters ensures equal opportunity, aligning with positive rights. These features can act as proxies for stability and models trained without these features can demonstrate a more balanced claim-approval rate across different income brackets.

On the legal rules around protected features [15], we decided to not consider gender for our model, to ensure no discrimination against a certain gender. Additionally, we will also be excluding

marital status, number of children, whether the driver is a single parent, education, occupation and type of car. This ensures freedom from direct discrimination, aligning with the negative rights theory. Because even when “neutral”, these features can act as proxies for the protected status and introduce bias in the system. Removing these features will significantly reduce discriminatory impact, although it might come at some cost to accuracy. We acknowledge this loss in accuracy when such features of high importance are excluded, but the utilitarianism approach allows us to balance the utility to the possible harm that could come from not doing so.

We have, however, also decided to retain the features of age and whether the driver’s license has been revoked to respect autonomy according to the Kantian school of thought. Although age is a protected feature, in our case it can be justified as a business necessity. It is a legally relevant factor for risk assessment, as drivers under the age of 25 are statistically more likely to get into accidents [16], which is why companies like Hertz charge them higher insurance [17]. As for number of licenses revoked, although it is a small dataset for whom the value is positive, it is highly relevant to driving behavior and public safety, ensuring the model reflects real-life driving concerns.

Potential Objections & Future Solutions

Our approach to preventing discrimination through feature selection raises several important concerns. Removing high-importance features like income and car type may significantly reduce model accuracy. As Barocas and Selbst note, "removing information may sometimes harm the very groups we seek to protect" [18]. To address this, we propose a balanced approach: using ethically neutral features for baseline predictions while incorporating sensitive features only with strict fairness constraints to maintain prediction quality without introducing bias.

The transparency provided by XAI methods, while beneficial, risks oversimplifying complex decisions and potentially revealing sensitive information. We recommend implementing tiered explanations tailored to different stakeholders, with all explanations undergoing fairness audits to prevent revealing protected characteristics.

Removing multiple features (home value, sex, income, car type, etc.) may create an underspecified model that fails to capture legitimate risk factors. This could lead to underfitting and unfair treatment of low-risk individuals within high-risk groups. We propose developing fairness-aware composite features that capture relevant risk information without encoding protected characteristics [19].

Traditional insurance risk assessment relies heavily on socioeconomic factors, creating tension between ethical ideals and business requirements. We recommend gradual implementation with parallel testing to quantify the "cost of fairness" against benefits in customer satisfaction and reduced legal risk.

Finally, counterfactual explanations might suggest impossible changes to customers, creating frustration rather than empowerment. We will implement feasibility constraints in our DiCE implementation, ensuring explanations focus on actionable changes within an individual's control [20].

By addressing these objections with thoughtful solutions, we balance non-discrimination imperatives with practical business considerations, recognizing that our ethical choices involve necessary trade-offs in the pursuit of fair algorithmic decision-making.

4. Disparate Impact Mitigation Report

Data & Code Overview

Our implementation evaluated and mitigated unfair bias in our insurance claim prediction model using Microsoft's FairLearn package. After excluding sensitive features like gender and income, we measured fairness using Demographic Parity and Equalized Odds metrics, finding minimal but present disparities (0.0017 and 0.0046 respectively) with an initial accuracy of 0.7467. Our mitigation techniques successfully reduced these disparities while maintaining reasonable performance, with only a slight decrease in accuracy to 0.7443, effectively balancing fairness with model utility.

Overview of Fairness Metrics & Mitigation

Predictive models are considered best when their predictions are accurate. But what if this accuracy comes at the cost of bias? Therefore, it is essential to not only assess the overall performance but to also ensure that the results are not disparate towards certain groups. This becomes more critical when we deal with sensitive, protected features like gender and age - features that have been historically subjected to bias.

Bias in predictive models can be evaluated via fairness metrics that quantify how equitably a model performs. It helps answer questions like “Are error rates, benefits and risks distributed fairly?” By pinpointing specific inequities, these metrics guide developers in selecting and tailoring interventions like adjusting threshold values and reweighting.

The recall parity metric, also called Equality of Opportunity, focuses on achieving the same true positive rate across groups, ensuring that individuals who deserve a positive outcome are equally likely to be identified by the model irrespective of their background. The false discovery rate metric ensures that the rate of false positives - instances when the model incorrectly predicts a positive outcome - is similar across groups. The equal odds metric certifies that the error rates, both false positives and false negatives, should be similar for all groups, thus balancing the risk of misclassification regardless of group membership while the demographic parity metric requires that the probability of favorable outcomes be equal across all demographic distributions.

Introducing these metrics is essential to abide by transparency, accountability and the rights of individuals regarding automated decision-making, as stated by GDPR [21]. With the increasing use of machine learning in decision-making, regulators are also beginning to focus specifically on algorithmic fairness [22]. Understanding and quantifying these metrics allows us to detect potential biases. Once these disparities are identified, fairness mitigation techniques can be applied. This process helps ensure that sensitive groups are not unduly harmed by automated decisions. However, addressing fairness often involves trade-offs with other model performance metrics like precision or overall accuracy. Awareness of these metrics helps in making informed decisions about where and how much to balance these trade-offs. From an ethical perspective, fair treatment of individuals is a core value in responsible data science practices.

Our Standard

For our model, we have opted to use demographic parity and equalized parity metrics. These metrics have been chosen by our intent to balance the equitable distribution of favorable outcomes across groups and to ensure that the model's error rates are not biased towards a particular section.

Demographic parity ensures that the proportion of favorable outcomes is consistent across groups, promoting a balanced treatment and minimizing disparate impact. Equalized odds, which require similar error rates across groups, are particularly important in maintaining equitable treatment in model performance. Our model demonstrated low scores to begin with, which was slightly increased post-mitigation, showcasing overall fairness preservation. We opted to not apply the recall metric as it is the proportion of actual positives correctly identified by the model. By ensuring Equalized Odds, we are implicitly ensuring that the model's ability to identify positive cases is balanced across groups.

Despite the fairness focused adjustments, the overall accuracy declined only slightly - 0.7467 to 0.7443 - demonstrating improvements in ensuring equity with minimal loss in performance. With metrics increasing slightly post-mitigation and accuracy decreasing slightly, we choose mitigation because prioritizing fairness and reducing disparate impacts is essential, even if it incurs a slight penalty.

Potential Objections & Future Solutions

While our approach to fairness mitigation shows promise, several important objections warrant consideration. First, the accuracy-fairness trade-off (0.7467 to 0.7443) raises questions about the business viability of fairness interventions. To address this, we propose implementing a sliding-scale approach where fairness constraints can be adjusted based on specific use cases and regulatory requirements, potentially using threshold optimization techniques to find optimal balance points as referenced in our class notes.

Some stakeholders might argue that our chosen metrics (Demographic Parity and Equalized Odds) may not fully capture all dimensions of fairness in insurance contexts. For instance, we did not implement recall parity, which could be particularly relevant for ensuring equal opportunity across groups. Future iterations could incorporate a multi-metric evaluation framework that balances different fairness definitions based on specific business and ethical priorities, like the approach suggested in the EU AI Act [22].

The fairness metrics we've implemented treat protected attributes as binary categories, potentially oversimplifying the complex, intersectional nature of discrimination. A person's identity encompasses multiple dimensions simultaneously (age, gender, socioeconomic status), and bias may manifest differently at these intersections. We recommend implementing intersectional fairness analysis in future versions to detect and mitigate bias that affects specific subgroups defined by multiple attributes [23].

Critics might contend that our post-processing approach to bias mitigation, while effective, doesn't address underlying data representation issues. To counter this, we propose exploring pre-processing techniques that address biased data representation before model training, potentially through reweighting or sampling methods that ensure balanced representation across protected groups.

Finally, the relatively small improvement in fairness metrics post-mitigation raises questions about whether more aggressive interventions are warranted. We suggest conducting a cost-benefit analysis that quantifies the utility gain for protected groups against overall performance reduction, adopting a utilitarian framework like what was discussed in our ethics overview. This would help determine appropriate thresholds for fairness intervention based on measurable impact rather than arbitrary statistical targets.

5. Safety & Liability Report

Data & Code Overview

Our implementation creates a secure, ethical AI assistant (InsuranceGPT) using Meta's Llama 3.1 model (8B) in LM Studio. The system prompt establishes ethical guardrails, permissible risk factors, and transparent calculation methodologies. The premium calculation uses a sequential multiplier approach starting with a base premium from vehicle value (BLUEBOOK), then applies risk multipliers for legally permissible factors including age, vehicle usage, car age, claim history, license status, MVR points, location, and commute time.

Each calculation is accompanied by transparent explanations breaking down how factors influenced the final premium, rounded to the nearest $25. Our testing confirmed the system successfully rejected problematic requests related to race/ethnicity adjustments, accessing others' personal information, and assistance with fraudulent activities, validating our ethical guardrail implementation. All calculation steps are documented to ensure transparency and auditability of recommendations.

Overview of LLMs for Safety and Vulnerability

LLMs introduce unique safety concerns in high-stakes domains like insurance, where recommendations directly impact financial decisions. Our analysis using ALERT and SALAD-Bench benchmarks identified several vulnerability categories.

Direct harm potential manifests through misinformation risk, where inaccurate calculations could cause financial harm. Despite explicit programming to avoid discrimination, latent biases may establish correlations between permitted variables and protected attributes, potentially resulting in indirect discrimination [24]. This concern aligns with research showing LLMs can perpetuate societal biases from their training data.

Adversarial vulnerabilities represent significant concerns, including prompt injection attempts to circumvent ethical guidelines (e.g., requests to adjust premiums based on race), privacy exploitation attempts, and requests for assistance with fraudulent activities. These vulnerabilities persist despite explicit safeguards [25].

Misuse potential extends beyond direct attacks to more subtle exploitation paths, including soliciting harmful advice and multi-step interactions where benign requests establish context for subsequent problematic queries. These findings reflect broader patterns where contextual manipulation often proves more effective than direct attacks [26].

Our Standard

InsuranceGPT upholds fairness, security, and privacy through three main principles: ethical guardrails, attack resistance, and privacy protection. It calculates premiums using only legally permissible features, explicitly rejecting discriminatory factors like gender, income, and family status. Instead, it focuses on objective risk factors such as vehicle value and driving behavior,

ensuring transparency and consistency in pricing. Customers receive clear explanations of how each factor influences their premium, promoting fairness throughout the process.

The system features input validation to block unethical requests, such as race-based premium adjustments or fraudulent activity. Ethical response protocols reject such requests while offering appropriate alternatives. Testing confirmed that the model can resist manipulation and maintain ethical integrity in its operations. To ensure privacy protection, InsuranceGPT follows data minimization principles, only collecting the necessary information for premium calculation. The system complies with GDPR [27], safeguarding personal data and protecting against social engineering attempts.

Premiums are determined by the vehicle's BLUEBOOK value, ensuring pricing is based on objective metrics rather than irrelevant characteristics. Risk multipliers are applied based on age, vehicle usage, claims history, and other factors, with clear reasoning for each adjustment. These multipliers follow principles of direct harm (e.g., higher risk for young drivers) and indirect harm (e.g., increased risk for commercial vehicles).

By excluding prohibited factors like gender and income, InsuranceGPT ensures fair treatment for all customers, maintaining both ethical fairness and legal compliance. Testing with ALERT and SALAD-Bench showed that the system balances utility with safety, ensuring ethical and legal standards are met.

Potential Objections and Future Solutions

Despite robust implementation, several limitations warrant consideration. The proxy variable challenge represents a significant concern, as permitted variables (particularly location) might serve as unintentional proxies for protected attributes. To address this, we propose implementing regular fairness audits that systematically evaluate premium recommendations across demographic groups.

Model drift presents another challenge, as underlying LLM behavior changes with updates. Safety measures effective today might require recalibration as models evolve. We recommend continuous monitoring protocols that regularly assess safety performance against benchmark datasets [28].

Advanced prompting techniques represent an evolving threat, as users develop new methods to circumvent safety measures. Our implementation resists known attacks, but new techniques will emerge. To counter this, we propose establishing an adversarial testing program that continuously evaluates the system against evolving attack methods.

Finally, our current system lacks comprehensive intersectional fairness analysis. We propose implementing methods to detect and mitigate bias affecting specific subgroups defined by multiple attributes, aligning with approaches described by Kearns et al. for subgroup fairness analysis [29].

By addressing these limitations through ongoing refinement, InsuranceGPT can provide valuable, personalized insurance recommendations while upholding ethical standards for fairness, safety, and accountability.

REFERENCE:

[1] California Consumer Privacy Act, California Civil Code Section 1798.100 et seq., 2018.

[2] California Privacy Rights Act, California Civil Code Section 1798.100 et seq., 2020.

[3] California Privacy Rights Act, California Civil Code Section 1798.120, 2020.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Recital 32.

[5] California Consumer Privacy Act, California Civil Code Section 1798.100(b), 2018.

[6] California Privacy Rights Act, California Civil Code Section 1798.121, 2020.

[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Recital 32

[8] Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq. (1974), Pub. L. 93-495, Title V, § 502, Oct. 28, 1974, 88 Stat. 1521.

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Articles 5 and 25, 2016.

[10] California Consumer Privacy Act, California Civil Code Section 1798.100(b), 2018.

[11] California Consumer Privacy Act, California Civil Code Section 1798.105 (Right to Delete), 2018.

[12] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Recital 71, 2016.

[13] Nozick, R. (1974). Anarchy, State, and Utopia. Basic Books.

[14] Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1).

[15] DeLucia, S.-A. (2023, May 19). Evaluating model fairness. Arize AI. https://arize.com/blog/evaluating-model-fairness/#:~:text=Let's%20dissect%20each.-,Sensitive%20Group%20Bias, address%20these%20concerns%20as%20well.

[16] Tefft, B. C. (2018, June 14). Rates of motor vehicle crashes, injuries and deaths in relation to driver age, United States, 2014-2015 - AAA Foundation for Traffic Safety. AAA Foundation for Traffic Safety - . https://aaafoundation.org/rates-motor-vehicle-crashes-injuries-deaths-relation-driver-age-united-states-2014-2015/

[17] Hertz. Car Rental: Save More on Rental Cars, Vans & Trucks. (n.d.). https://www.hertz.com/rentacar/reservation/

[18] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671-732.

[19] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? CHI Conference on Human Factors in Computing Systems Proceedings.

[20] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2).

[21] *Legal text*. General Data Protection Regulation (GDPR). (2024, April 22). https://gdpr-info.eu/

[22] *EU AI act: First regulation on artificial intelligence: Topics: European parliament*. Topics | European Parliament. (n.d.). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[23] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. Proceedings of the 35th International Conference on Machine Learning.

[24] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

[25] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Le, Q. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

[26] Perez, E., Huang, S., Song, H. F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

[27] Legal text. General Data Protection Regulation (GDPR). (2024). https://gdpr-info.eu/

[28] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

[29] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. Proceedings of the 35th International Conference on Machine Learning.