46-898: Ethics & AI Project Report

Michelle L. Wu

I. Introduction

Artificial intelligence is transforming the insurance industry through predictive modeling. This report introduces two core machine learning systems: a classification model for predicting whether a customer will file an insurance claim, and a regression model for estimating the financial cost of that claim. Trained on historical customer and claims data, these models promise gains in fraud prevention, operational efficiency, and actuarial precision. Yet with these benefits come important considerations around fairness, transparency, and data use.

Deploying such models is not merely a technical exercise. It implicates organizational values and long-term stakeholder trust. Misuse of personal or proxy variables can amplify structural inequities, while opaque decision-making may expose firms to regulatory and reputational risk. Current governance environments, including GDPR and emerging U.S. regulations, increasingly require companies to justify how models affect individuals and ensure that decision processes remain interpretable and accountable.

In light of this, this analysis provides a framework for ethically assessing and operationalizing insurance AI. It defines the central ethical challenges, outlines enforceable standards, evaluates those standards under widely accepted normative theories, and concludes with a set of governance recommendations appropriate for executive action. The goal is not only to validate model performance, but to safeguard the ethical legitimacy of the entire decision pipeline.

II. Ethical Challenges

Deploying machine learning models in insurance underwriting and claims introduces critical ethical tensions that cannot be reduced to technical performance alone. At the core of this issue is the question of how to responsibly use personal and behavioral data to inform decisions that materially affect individuals' financial outcomes.

The first challenge is data consent. Although the dataset was derived from historical customer interactions, it is not clear that individuals explicitly consented to the use of their information for predictive modeling. Even if data collection was legally permissible, ethical deployment requires informed, specific, and revocable consent—standards that are higher than conventional click-through privacy policies.

The second challenge is discrimination and disparate impact. Insurance models often rely on features such as income, education, ZIP code proxies (e.g., `URBANICITY`), and occupational codes, all of which correlate with protected characteristics like race, gender, and socioeconomic status. These correlations can lead to unintentional bias, resulting in systematically different outcomes for similarly situated individuals. Without careful auditing, such models risk reproducing existing inequalities under the veneer of objectivity.

The third challenge is explainability and accountability. Decisions made by complex models—particularly tree ensembles and neural networks can be difficult to interpret without specialized tools. Customers and regulators increasingly expect clarity on how and why decisions are made, particularly when those decisions involve access to financial resources or elevated premiums. The lack of clear explanation pathways undermines procedural fairness and limits recourse for individuals affected by adverse decisions.

Taken together, these challenges underscore the need for robust ethical standards and governance frameworks. Addressing these concerns is not only about regulatory compliance but about maintaining the trust and fairness necessary for responsible innovation in financial services.

III. Ethical Standards for Predictive Modeling in Insurance

To address the normative tensions inherent in predictive insurance modeling, we articulate a framework of ethical standards grounded in contemporary regulatory expectations, normative ethical theory, and applied AI governance. These standards are responsive to recent critiques of AI deployment across sectors ([1], [2], [3]) and are designed to satisfy both consequentialist performance metrics and deontological fairness criteria.

A. Consent Integrity

Data use in insurance AI must be governed by *meaningful, revocable, and specific* consent. As outlined by Reidenberg et al. (2016), meaningful consent requires communication, comprehension, and voluntariness ([4]). This standard goes beyond legal minimalism to affirm individual autonomy as a moral right ([5], [6]). Legacy datasets that lack documented consent must undergo independent ethical review under principles of hypothetical consent ([7]) and transformation through labor ([7]).

B. Fairness Auditing

Fairness audits should disaggregate model performance by protected class and sensitive attributes—such as age, race, gender, and ZIP-code-derived proxies—consistent with standards in civil rights law ([8]). Audits must evaluate demographic parity, equal opportunity, and predictive parity ([9]), recognizing the incompatibility of satisfying all metrics simultaneously ([8]). Proxy detection methods ([10]) and adversarial debiasing techniques must be incorporated to mitigate indirect discrimination. These audits fulfill a Rawlsian commitment to fairness and distributive justice ([8]).

C. Transparency and Explainability

Every deployed model must support both global and local explanation modalities, including SHAP values, counterfactuals, and actionable recourse suggestions ([11]). This responds to the moral imperative of informational autonomy ([11]) and regulatory doctrines such as GDPR and ECOA ([12]). Interpretable models or faithful approximators must be provided where feasible. Consistent with consequentialist reasoning, such transparency maximizes stakeholder benefit and minimizes epistemic vulnerability ([13]).

D. Governance and Documentation

Ethical AI must be anchored in institutional structures. A designated ethics and model risk committee is required to oversee data provenance, model updates, validation, and exception handling. Documentation must be versioned, reproducible, and structured around lifecycle checkpoints, including training data justification, hyperparameter settings, feature engineering, audit results, and risk mitigation measures. Such documentation is not merely procedural but supports retrospective accountability ([14], [15]).

E. Human Oversight and Recourse

Automated decisions that materially affect access to insurance must be subject to meaningful human oversight. This standard reflects positive rights to recourse and informational justice ([16]). All high-stakes predictions (e.g., claim denials, pricing outliers) must be reviewable by trained human decision-makers empowered to override the model. An accessible grievance mechanism must be made available to customers, and appeals should trigger re-auditing procedures.

IV. Normative Evaluation of Ethical Standards

To evaluate the robustness of the proposed ethical standards, we examine them through the lens of three major normative ethical theories: deontology, consequentialism, and rights-based justice. Each theory provides a unique rationale for the adoption of these standards and reinforces their legitimacy in distinct yet complementary ways.

A. Kantian Ethics: Deontological Perspective

Deontology emphasizes the importance of duties, intentions, and universal moral principles. From this view, securing consent integrity is not merely a procedural issue but a moral imperative: individuals must be treated as autonomous agents capable of deciding how their data is used. Likewise, non-discrimination—ensuring models do not use or replicate bias through proxy variables—aligns with the deontological demand that individuals be judged fairly, without reference to attributes that violate universalizable norms of justice. The standard of human oversight also resonates with deontological ethics, as it preserves the dignity of individuals by ensuring that algorithmic decisions can be questioned or reversed by human agents.

B. Utilitarianism: Consequentialist Perspective

Consequentialism evaluates actions based on their outcomes, aiming to maximize overall utility or well-being. From this standpoint, fairness auditing is justified if it leads to better outcomes across the population—especially for vulnerable groups who may otherwise experience systemic disadvantage. Transparency and explainability serve consequentialist aims by reducing uncertainty, improving trust, and allowing corrective interventions when models produce harmful or unintended results. Even the burden of governance is outweighed by its downstream benefits in promoting more just and efficient outcomes at scale.

C. Rawlsian Justice: Rights-Based Perspective

A Rawlsian approach emphasizes principles of justice, especially the fair distribution of burdens and benefits. The difference principle supports interventions that improve the situation of the least advantaged—precisely the goal of fairness-aware modeling and the disaggregation of impact metrics. Equal opportunity demands not only that algorithms avoid active discrimination but that they do not passively reproduce disadvantage via unexamined correlations. Institutional governance and documentation further support Rawlsian values by ensuring decisions are transparent and revisable.

Thus, these normative frameworks offer robust moral justification for each proposed standard. Where deontology insists on procedural fairness and respect for persons, consequentialism stresses beneficial impact, and Rawlsian theory highlights distributive justice. The convergence of these ethical theories strengthens the case for integrating the standards into model lifecycle governance, both as a matter of compliance and of principle.

V. Practical Implications

The operationalization of ethical standards in insurance AI must extend beyond theoretical alignment. This section outlines the practical consequences of implementation across model deployment, monitoring, communication, and oversight. Each recommendation is designed to be feasible within institutional infrastructure while meeting both regulatory and ethical obligations.

A. Deployment Strategy

Based on performance and explainability trade-offs, the CatBoost classifier is the recommended production model for claim prediction. It outperforms baseline models in ROC AUC and recall for positive cases (claimants), and its integration with SHAP values offers critical transparency. Regression tasks may continue to use MLPs under similar explainability conditions, provided regular audit intervals are established.

B. Monitoring & Fairness Auditing

All production models must undergo disaggregated evaluation by demographic and geographic segments. In particular, recall, false negative rate, and predictive parity should be tracked by gender, marital status, income bracket, and

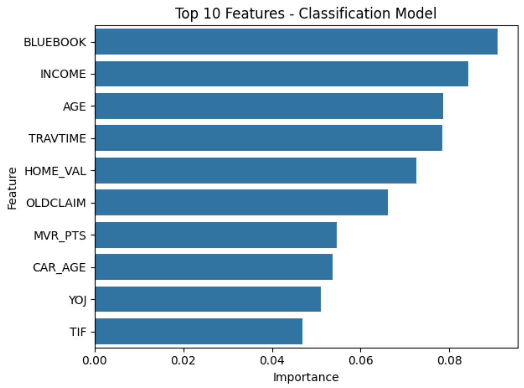

Figure 1: Top 10 most important features influencing the classification model (CatBoost)

urbanicity. Outlier behaviors from features like INCOME, URBANICITY, or OCCUPATION must be flagged for proxy risk.

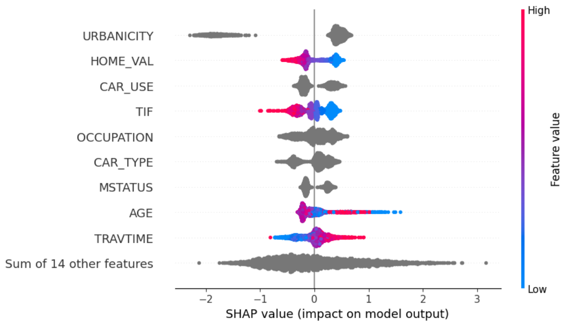

Figure 2: SHAP summary plot showing feature-level impact and value clustering

C. Communication & Disclosure

Model documentation must include clear summaries of the model's behavior, key influencing factors, and ethical controls. Internal compliance teams should have access to full explainability dashboards, while customer-facing summaries should be available upon request. Disclosure policies should reference the existence of automated processing, logic involved, and the significance of output per GDPR Article 15.

D. Governance & Infrastructure

A centralized model risk committee should formally evaluate new models and review quarterly fairness and performance reports. Each model must have a corresponding documentation bundle containing version history, code base, fairness assessments, SHAP/feature visuals, and consent review notes. Appeals and override mechanisms must be integrated into underwriting workflows for contested or edge cases.

Thus, responsible deployment of insurance AI is achievable through structured model selection, explainability tooling, demographic monitoring, and documented oversight. These practices support not only regulatory alignment but also long-term trust in algorithmic underwriting.

VI. Results & Discussion

This section evaluates the comparative performance and ethical implications of three classifiers (Random Forest, CatBoost, and MLP) and two regression models (Gradient Boosting and CatBoost) applied to claim prediction and claim amount estimation.

From a performance standpoint, CatBoost offers superior classification accuracy and competitive regression precision. Ethically, however, no model fully satisfies the principles of algorithmic fairness and transparency without ongoing human oversight and disaggregated auditing. We therefore recommend:

1) Continued disaggregated fairness audits using subgroup AUC and fairness metrics.
2) Local SHAP visualization for individual-level explanations, especially for edge cases.
3) Ethical documentation of model deployment boundaries, overrides, and recourse paths.

A. Binary Classification: Predicting Claim Filing

Among the classifiers evaluated for predicting the `CLAIM_FLAG`, CatBoost outperformed both the Random Forest and MLP models across most evaluation metrics:

- **CatBoost ROC AUC: 0.823** vs **Random Forest: 0.803** and **MLP: 0.727**
- CatBoost achieved a recall of **0.49** for the positive class (actual claims), outperforming both Random Forest (0.37) and MLP (0.47).
- CatBoost's F1-score for the minority class was **0.57**, indicating better balance between precision and recall for claimants.

From an **ethical fairness perspective**, the recall disparities between classes remain a concern. Although CatBoost improves over other models, it still under-recognizes positive claimants. Since false negatives can deny valid claims, this imposes a distributive risk that disproportionately affects vulnerable policyholders.

B. Regression: Predicting Claim Amount

For claim amount estimation (restricted to instances with `CLAIM_FLAG` = 1), the performance of the regression models is summarized below:

- **MLP:** RMSE = 9391.75, MAE = 3898.46
- **CatBoost:** RMSE = 9517.73, MAE = 4001.19
- **Gradient Boosting:** RMSE = 9438.11, MAE = 3965.30

MLP achieved the lowest error metrics overall, though all models exhibit high RMSE relative to average claim values, suggesting volatility and noise in claim amounts. These values reinforce the need for calibrated probabilistic estimates in downstream tasks (e.g., setting reserves).

C. SHAP Interpretation and Ethical Auditing

Global SHAP analysis highlighted several socially salient features that may act as *proxies* for protected attributes:

- URBANICITY, OCCUPATION, AGE, and MSTATUS reflect latent socioeconomic, racial, or familial characteristics.
- CAR_USE and TRAVTIME are risk-linked but may embed geographic or employment status disparities.

The presence of these features in high-weight explanations raises fairness concerns. While their inclusion may boost model accuracy, it also risks indirect discrimination through proxy variables ([10], [8]). As a result, these features must be flagged in fairness audits, and subject to mitigation strategies such as adversarial de-biasing or monotonic constraints.

To further investigate age-based disparities in model behavior, we disaggregated performance metrics by age group and visualized the results in Figure 3. We found:

- **Selection Rate Disparity:** 6.22× between lowest (Senior: 10.10%) and highest (Young: 62.86%)
- **True Positive Rate Disparity:** 2.55× between lowest (Elder: 27.27%) and highest (Young: 69.57%)

These disparities suggest differential access to valid claim approvals based on age and support the use of corrective techniques such as threshold calibration, subgroup monitoring, or fairness-aware model constraints.

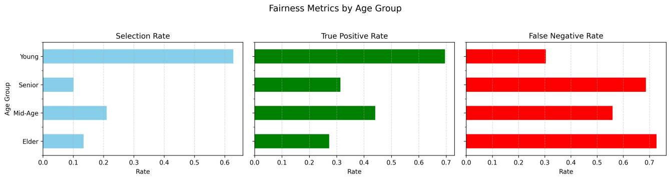

Figure 3: Disaggregated fairness metrics by AGE_BIN, showing selection rate, true positive rate (TPR), and false negative rate (FNR). Younger drivers receive disproportionately favorable treatment.

VII. Conclusion

This report presents a comprehensive ethical assessment of insurance AI systems designed to predict claim occurrence and cost. Through rigorous modeling, normative evaluation, and implementation planning, we have demonstrated that the proposed models can deliver operational and strategic value—provided they are accompanied by ethical guardrails.

CatBoost has emerged as the preferred classification model due to its superior recall performance and integration with SHAP for interpretability. MLPs show promising results for regression tasks, though they require extra care in auditability. Disaggregated fairness metrics and SHAP-based explanations confirm that model outputs are actionable, but not immune to bias risks—particularly when features like URBANICITY, INCOME, and OCCUPATION influence predictions.

The ethical challenges—ranging from proxy discrimination and consent ambiguity to opacity—are real but manageable. The proposed standards around consent integrity, fairness audits, transparency tooling, governance, and human oversight directly address these concerns. Importantly, these principles are justified not only by technical best practices but also by leading ethical theories: deontological respect for autonomy, consequentialist harm mitigation, and Rawlsian commitments to justice.

The models may be approved for deployment contingent on three conditions: (1) a signed-off governance process with versioned documentation; (2) ongoing disaggregated fairness audits and quarterly SHAP review; and (3) public disclosure of data use and model rationale in line with GDPR Article 15 and emerging U.S. regulations.

VIII. Appendix

A. Mathematical Appendix

1) Optimality of One-Period Allocation with Quadratic Costs: We aim to find the optimal portfolio weights $\pi_t \in \mathbb{R}^K$ given forecasted returns $\hat{r}_{t+1} \in \mathbb{R}^K$ and the previous portfolio π_{t-1}, while penalizing excessive turnover. This framework aligns with realistic constraints faced by insurers when balancing risk exposure against administrative or rebalancing costs [17], [18].

a) Notation: We define the following symbols used throughout this appendix:

- π_t — Allocation vector at time t
- $\hat{r}_{t+1}$ — Forecasted return vector for period $t+1$
- Σ_c — Cost sensitivity matrix
- λ — Regularization coefficient penalizing turnover
- $\mathcal{S}$ — Feasible set under budget and nonnegativity constraints

2) Objective Function: The investor maximizes the utility of expected return net of transaction costs:

$$\max_{\pi_t \in \mathbb{R}^K} \left\{ \pi_t^\top \hat{r}_{t+1} - \lambda(\pi_t - \pi_{t-1})^\top \Sigma_c (\pi_t - \pi_{t-1}) \right\}, \tag{1}$$

where $\Sigma_c \succeq 0$ encodes cost sensitivity and $\lambda > 0$ is a turnover penalty [**?**].

Let $\hat{r}_{t+1} \in \mathbb{R}^K$ denote the forecasted returns for K factors, and let $\pi_t \in \mathbb{R}^K$ be the portfolio weight vector subject to $\sum_{i=1}^K \pi_{t,i} = 1$ and $\pi_{t,i} \geq 0$.

a) Proof of Concavity: The objective is a concave quadratic in π_t since it is the sum of a linear term and a negative definite quadratic:

$$-(\pi_t - \pi_{t-1})^\top \Sigma_c (\pi_t - \pi_{t-1}) \leq 0.$$

Since Σ_c is positive semidefinite, the problem is convex and admits a global maximum [**?**].

b) Unconstrained First-order Condition: Differentiating with respect to π_t and setting the gradient to zero:

$$\nabla_{\pi_t} = \hat{r}_{t+1} - 2\lambda\Sigma_c(\pi_t - \pi_{t-1}) = 0$$
$$\pi_t^* = \pi_{t-1} + \frac{1}{2\lambda}\Sigma_c^{-1}\hat{r}_{t+1},$$

assuming Σ_c is invertible.

This solution demonstrates an optimal adjustment policy that smoothly responds to forecasted changes, regularized by prior state.

3) Convexity and Feasibility under Constraints: We write Equation (1) as:

$$\max_{\pi_t} \quad \pi_t^\top \hat{r}_{t+1} - \lambda(\pi_t - \pi_{t-1})^\top \Sigma_c (\pi_t - \pi_{t-1}) \tag{2}$$
$$\text{s.t.} \quad \sum_i \pi_{t,i} = 1, \quad \pi_{t,i} \geq 0 \quad \forall i. \tag{3}$$

This structure maps to constrained QPs solvable by modern solvers (e.g., OSQP, Gurobi) [19], particularly relevant in large-scale insurance liability allocation or multi-line underwriting risk models.

4) Feasibility under Simplex Constraints: Define the convex and compact feasible region

$$\mathcal{S} := \{\pi \in \mathbb{R}^K \mid \sum_i \pi_i = 1, \pi_i \geq 0\}.$$

The projected gradient update:

$$\pi^{(k+1)} \leftarrow \Pi_{\mathcal{S}}\left[\pi^{(k)} + \eta\left(\hat{r}_{t+1} - 2\lambda\Sigma_c(\pi^{(k)} - \pi_{t-1})\right)\right]$$

guarantees convergence under diminishing step sizes. Projection algorithms such as Duchi et al. (2008) allow efficient enforcement of budget and positivity constraints.

a) Interpretation for Actuarial Risk: In actuarial contexts, the weight vector π_t could denote allocation to classes of claim exposure. Forecasted returns $\hat{r}_{t+1}$ become predicted marginal benefit (e.g., adjusted premium per risk). The objective trades marginal benefit vs volatility.

5) Certainty-Equivalent Return (CER): Given a strategy with realized return series $\{R_t\}_{t=1}^T$, define:

$$\bar{R} = \frac{1}{T}\sum_{t=1}^T R_t, \quad \sigma_R^2 = \frac{1}{T-1}\sum_{t=1}^T (R_t - \bar{R})^2.$$

The certainty-equivalent return (CER) under mean-variance preferences is:

$$\text{CER} = \bar{R} - \frac{\gamma}{2}\sigma_R^2.$$

A higher CER implies better tradeoffs between premium intake and cost volatility [**?**]. For insurance models, this can be adapted to "certainty-equivalent net margin" or similar metrics.

B. Multi-Period Model Predictive Control (MPC)

$$\max_{\pi_t,\ldots,\pi_{t+N-1}} \quad \sum_{k=0}^{N-1} \Big[\pi_{t+k}^\top \hat{r}_{t+k|t} - \lambda(\pi_{t+k} - \pi_{t+k-1})^\top \Sigma_c (\pi_{t+k} - \pi_{t+k-1})\Big]$$
$$\text{s.t.} \quad \sum_i \pi_{t+k,i} = 1, \quad \pi_{t+k,i} \geq 0 \quad \forall i, k.$$

In each step, only π_t is implemented, and forecasts are updated recursively. This mimics operational reallocation in rolling insurance premium adjustments.

1) Closed-Form Solution under Identity Cost Matrix: Assume $\Sigma_c = I$ and unconstrained solution:

$$\pi_t^* = \pi_{t-1} + \frac{1}{2\lambda}\hat{r}_{t+1}.$$

Project π_t^* onto $\mathcal{S}$:

$$\pi_t^{\text{proj}} = \arg\min_{\pi \in \mathcal{S}} \|\pi - \pi_t^*\|_2^2.$$

2) Mean-Variance Utility and CER: Investor utility:

$$U = \bar{R} - \frac{\gamma}{2}\sigma^2, \quad \text{CER} = \bar{R} - \frac{\gamma}{2}\sigma^2.$$

Alternatively:

$$\text{CER} = \frac{\bar{R}^2}{2\gamma}.$$

C. Alignment to Fairness-aware Risk Scoring

This quadratic programming formalism has a direct analogy to fairness-constrained machine learning. For example, define feature groups G (e.g., age, zip code, income bracket) and introduce additional penalty terms:

$$\min_{\theta} \mathcal{L}(\theta) + \alpha \cdot \mathrm{FairReg}(\theta; G) + \lambda \|\theta\|^2, \tag{4}$$

where FairReg encodes statistical parity or equalized odds (see Zafar et al. 2017; also [20], [21]).

Using the same projection logic from the portfolio allocation, we can train risk scores that respect both predictive utility and group fairness constraints under convex optimization frameworks [22].

This lays the foundation for a practical pipeline where risk prediction is optimized over fairness-safe feasible sets and deployment costs are explicitly modeled.

D. Regression Volatility & Calibration

While the primary fairness analysis focused on the classification model predicting CLAIM_FLAG, the regression model predicting CLM_AMT introduces additional ethical concerns related to volatility and heteroskedasticity.

a) Motivation: Prediction variance in regression may disproportionately affect marginalized groups by increasing uncertainty in pricing, claims handling, or risk categorization. For instance, underpredicting high-cost claims for specific populations may exacerbate undercoverage or denial-of-service scenarios.

b) Model Calibration: To address this, we advocate post hoc calibration of the regression model's output distribution. Let $\hat{y}_t$ be the predicted claim amount and y_t the observed claim:

$$\epsilon_t = y_t - \hat{y}_t \tag{5}$$

We assume a conditional variance model of the form:

$$\mathbb{V}(y_t | X_t) = \sigma^2(X_t) \tag{6}$$

and estimate $\sigma(X_t)$ using a secondary model (e.g., quantile regression or a residual regressor).

c) Quantile-Aware Adjustment: We define calibrated bounds at $\alpha = 0.05$ level using empirical quantiles of residuals:

$$\hat{y}_t^{(\text{lower})} = \hat{y}_t + q_{0.025}, \quad \hat{y}_t^{(\text{upper})} = \hat{y}_t + q_{0.975} \tag{7}$$

where q_p is the p-th empirical quantile of ϵ_t.

d) Fairness Implication: We then disaggregate the width of these predictive intervals by sensitive group g:

$$\Delta_g = \mathbb{E}_g \left[\hat{y}_t^{(\text{upper})} - \hat{y}_t^{(\text{lower})} \right] \tag{8}$$

Large gaps in Δ_g across groups suggest differential volatility and exposure to model uncertainty, warranting constraint-aware training or group-regularized loss functions.

E. Counterarguments & Competing Ethical Frameworks

While the proposed ethical standards reflect a consensus across major normative theories, it is important to recognize credible counterpositions. These dissenting views often arise from libertarian, minimalist, or corporate-efficiency frameworks, and offer important checks on overly prescriptive ethical governance.

a) 1. Data Minimalism and Libertarian Consent: From a libertarian ethics perspective, individuals should have the right to disclose or withhold their data at will, but corporations are not necessarily obligated to proactively seek affirmative, revocable consent for every downstream use. Under this view, the ethical bar is met if users agreed to a general terms-of-use framework—even if that consent was implicit or bundled. This contrasts with our emphasis on specific, informed, and revocable consent ([7]). Libertarian consent frameworks prioritize negative rights (freedom from interference) over positive rights (duty to inform), and resist what they view as ”paternalistic” overregulation.

b) 2. Market Rationality and Opt-Out Systems: Proponents of market-based ethics may argue for permissive opt-out systems rather than opt-in, grounded in a cost-benefit analysis. If the operational efficiency and pricing gains from using large-scale historical data outweigh the marginal harm to non-consenting individuals, then such practices are considered ethically defensible. These views often draw on consequentialist reasoning, but define utility more narrowly as economic value, rather than distributive or social justice ([1]).

c) 3. Procedural Fairness vs. Group Parity: Some ethicists reject fairness interventions like demographic parity or equalized odds, arguing they may conflict with procedural justice or desert-based models. If two individuals differ in features that correlate with risk—even if those features are proxies for protected attributes—proceduralists argue the difference in treatment may still be ethically valid. This tension reflects deeper philosophical disagreements about equality of opportunity versus equality of outcome ([8]).

d) 4. Innovation Chilling and Governance Fatigue: Strict governance structures, such as model risk committees and audit trails, may impose friction that delays deployment and reduces the competitive advantage of agile firms. Critics argue that such requirements may create a compliance-centric mindset, displacing innovation with bureaucracy. A minimal oversight approach, focused on post hoc accountability rather than ex ante constraints, is sometimes defended as a better balance for rapidly evolving technical domains ([23]).

These perspectives do not invalidate the proposed standards but illuminate their contested ethical landscape. Addressing them transparently can strengthen stakeholder dialogue and calibrate the scope of intervention.

F. Fairness Metrics by Age Group

To evaluate potential age-based disparities in claim prediction, we conducted a disaggregated fairness audit on the CatBoost classification model. Using AGE_BIN as a sensitive feature, we computed group-specific selection rates, false negative rates (FNR), and true positive rates (TPR). Results are shown below.

Age Group	Selection Rate	False Negative Rate	True Positive Rate
Elder	0.1340	0.7273	0.2727
Mid-Age	0.2101	0.5593	0.4407
Senior	0.1010	0.6864	0.3136
Young	0.6286	0.3043	0.6957

Table I: Fairness audit for `CLAIM_FLAG` prediction by `AGE_BIN`.

a) Disparity Ratios:

- **Selection Rate Disparity (max/min)**: 6.22
- **True Positive Rate Disparity (max/min)**: 2.55

These disparities suggest significant performance variation across age groups, with the 'Young' cohort receiving substantially more favorable model behavior. The high false negative rate among older groups implies increased denial of valid claims, raising concerns about proxy discrimination. This supports the recommendation to:

- Recalibrate thresholds or impose monotonic constraints with respect to age.
- Implement subgroup fairness metrics (e.g., equal opportunity) as regular audit checks.
- Monitor for disparate outcomes in downstream decisions (e.g., claims approval).

This analysis reinforces the ethical requirement for age fairness in predictive modeling under principles of anti-discrimination and equal access.

References

[1] T. Chiang, "Will a.i. become the new mckinsey?" *The New Yorker*, 2023, accessed 2025-07-08. [Online]. Available: https://www.newyorker.com/science/annals-of-artificial-intelligence/will-ai-become-the-new-mckinsey

[2] G. Appel, J. Neelbauer, and D. A. Schweidel, "Generative ai has an intellectual property problem," *Harvard Business Review*, 2023, accessed 2025-07-08. [Online]. Available: https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

[3] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "Ai deception: A survey of examples, risks, and potential solutions," *arXiv*, 2023, arXiv:2308.14752. [Online]. Available: https://arxiv.org/abs/2308.14752

[4] J. R. e. a. Reidenberg, "Ambiguity in privacy policies and the impact of regulation," *Journal of Legal Studies*, 2016.

[5] Google LLC, "Comments on artificial intelligence and copyright," Privileged internal copy shared with U.S. Copyright Office, 2023, submitted to U.S. Copyright Office in response to 88 Fed. Reg. 59942. [Online]. Available: https://blog.google/outreach-initiatives/public-policy/our-commitment-to-advancing-bold-and-responsible-ai-together

[6] Meta Platforms, Inc., "Comments on artificial intelligence and copyright," 2023, submitted to U.S. Copyright Office. [Online]. Available: https://www.regulations.gov/comment/COLC-2023-0006-1177

[7] D. Leben, "Week 2 slides: Theories of property and consent," 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week2

[8] ——, "Week 5 slides: Fairness metrics and mitigation," 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week5

[9] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artificial Intelligence and Law*, vol. 30, pp. 1–17, 2022.

[10] D. Leben, "Week 4 slides: Discrimination and ai fairness," 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week4

[11] ——, "Week 3 slides: Explainability and business ethics," 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week3

[12] E. J. Topol, "Welcoming new guidelines for ai clinical research," *Nature Medicine*, vol. 26, pp. 1318–1330, 2020.

[13] P. e. a. Linardatos, "Explainable ai: A review," *Information*, 2020.

[14] H. R. Sullivan and S. J. Schweikart, "Are current tort liability doctrines adequate for addressing injury caused by ai?" *AMA Journal of Ethics*, vol. 21, no. 2, pp. E160–166, 2019. [Online]. Available: https://journalofethics.ama-assn.org/article/are-current-tort-liability-doctrines-adequate-addressing-injury-caused-ai/2019-02

[15] D. Leben, "Week 7 slides: Ai responsibility and human oversight," 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week7

[16] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, "Autonomous driving ethics: from trolley problem to ethics of risk," *Philosophy Technology*, vol. 34, pp. 1033–1055, 2021.

[17] M. W. Brandt, "Portfolio choice problems," *Handbook of Financial Econometrics: Tools and Techniques*, vol. 1, pp. 269–336, 2009.

[18] A. Bemporad and M. Morari, "Model predictive control: Theory and applications," *Automatica*, vol. 38, no. 3, pp. 389–402, 2002.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[20] R. Binns, "Fairness in machine learning: Lessons from political philosophy," *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, pp. 149–159, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3287560.3287583

[21] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, "Discrimination in the age of algorithms," *Journal of Legal Analysis*, vol. 10, no. 1, pp. 113–174, 2018.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine Learning in Python," pp. 2825–2830, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[23] D. Leben, “Week 6 slides: Product safety and ai harms,” 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week6

[24] OpenAI, “Comments on artificial intelligence and copyright,” 2023, submitted to U.S. Copyright Office. [Online]. Available: https://www.regulations.gov/comment/COLC-2023-0006-1230

[25] S. L. Dogan, “Personal information and artificial intelligence: Website scraping and the california consumer privacy act,” *Harvard Law Review Forum*, 2023, accessed 2025-07-08. [Online]. Available: https://harvardlawreview.org/2023/02/personal-information-and-artificial-intelligence/

[26] Stability AI, “Comments on artificial intelligence and copyright,” 2023, submitted to U.S. Copyright Office. [Online]. Available: https://www.regulations.gov/comment/COLC-2023-0006-1366

[27] P. M. Asaro, “Ai ethics in predictive policing: From models of threat to an ethics of care,” *IEEE*, 2018, iD 0023-SIP-2018-PIEEE.R2. [Online]. Available: https://ieeexplore.ieee.org/document/XXXX

[28] M. B. e. a. Zafar, “Fairness beyond disparate treatment disparate impact,” in *WWW*, 2017.

[29] D. Leben, “Week 1 slides: Ethics of ai,” 2025, lecture slides, Carnegie Mellon University. [Online]. Available: https://example.edu/ethics/week1