

Evaluating Human-AI Decision-Making Policies Under Imperfect Proxy Labels

Yi Hao Ming Wang^{1,*}, Luke Guerdan², Charvi Rastogi³ and Kenneth Holstein²

¹University of Southern California

²Carnegie Mellon University

³Google DeepMind

Abstract

AI systems are increasingly introduced to augment human decision-making in high stakes environments. However, measuring the performance of hybrid human-AI teams is challenging because the outcomes used for evaluation are often an imperfect proxy for the goals of human decision-makers. For example, while a doctor may wish to accurately diagnose and treat disease, diagnosis labels in medical records may imperfectly reflect patients' medical condition (e.g., due to limited health insurance). This gap between the goals of human decision-makers and the outcome observed in data makes it challenging to reliably measure the performance of alternative decision policies (e.g., human-only, AI-only, Human+AI). In this work, we develop evaluation tools to support robust comparison of decision policies under imperfect proxy labels. Our tools enable practitioners to assess whether the relative performance of different policies holds under plausible assumptions on the quality of proxy labels. Using our framework, we re-examine eight influential studies comparing human versus AI decisions across domains such as healthcare, lending, and child welfare. We find that the relative performance of decision policies can change under different assumptions about the gap between observed labels and the goals of human decision-makers. This work underscores the importance of developing robust evaluation approaches to evaluate the efficacy of different approaches designed to improve communication, coordination, and collaboration in human-AI teams.

Keywords

Evaluation, Human-Algorithm Decision-Making, Measurement, Validity

1. Introduction

AI systems are increasingly introduced under the rationale that they improve performance over an alternative (e.g., human-only) decision-making policy [1]. For example, AI systems have been developed with the goal of improving the accuracy and consistency of decisions in healthcare and education domains [2, 3]. When introducing an AI system to form a hybrid human-AI team, it is critical to examine the relative performance of alternative human-AI team configurations. We call a specific configuration of a human-AI team a *decision policy*. A common policy comparison involves evaluating human-only versus AI-only decisions to establish baseline performance (e.g., [4, 5]). A second frequently-conducted policy comparison involves evaluating whether a hybrid human-AI team produces better decisions than human-only or AI-only decisions alone — i.e., measuring *human-AI complementarity* [6, 7, 8].

However, reliably comparing decision policies is challenging because **the “ground-truth” labels used for evaluation often imperfectly reflect the goals of human decision-makers** [9]. For example, risk assessments used to inform judicial pre-trial release decisions target re-arrest as a proxy for re-offense [10], while clinical models used to inform program enrollment decisions target measures of healthcare utilization (e.g., cost) as a proxy for medical needs [11]. These proxies can be impacted by *target-construct misalignment* when they systematically differ from the unobservable construct of interest to humans (e.g., “medical needs”, “criminal risk”) [12, 9]. Critically, it is often impractical or impossible to obtain labels that reflect the full scope of decision-makers objectives, given that they often reflect an unobserved latent construct. **As a result, the validity of a policy comparisons hinges on the extent to which proxy labels reflect the broader objectives of humans.**

AutomationXP25: Hybrid Automation Experiences, April 27, 2025, Yokohama, Japan. In conjunction with ACM CHI'25.

*Corresponding author.



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Therefore, in this work, we develop a method for comparing decision-making policies in human-AI teams under imperfect proxy labels. Our framework allows practitioners to examine the sensitivity of policy comparisons to different assumptions about the quality of observed proxy labels. Our method consists of two specific evaluation strategies: *measurement sensitivity analysis* enables practitioners to assess the largest magnitude of measurement error permissible before the relative performance between policies is indeterminate. *Cost ratio curves* enables practitioners to examine the relative performance of decision policies under different costs of false positive and false negative decisions.

We leverage our methodology to re-analyze eight widely-cited human-AI decision-making policy comparisons reported in prior human-AI decision-making literature. We re-examine comparisons of (1) human-only versus AI-only policies [13? , 14, 15], (2) human+AI versus AI-only policies [16, 17, 11], and (3) an AI-only versus a checklist-based policy [18]. **Our analysis finds that performance comparisons are sensitive to outcome measurement error.** Of the studies included in our empirical application, all eight become inconclusive when 5% or more outcomes are mismeasured. This error tolerance is well below existing measurement error estimates for re-arrest outcomes carefully estimated in the criminology literature ($\approx 18\%$). This indicates that observed policy performance differences may be attributable to measurement error as opposed to meaningful differences in decision policy performance. Furthermore, we find that the optimal decision policy (e.g., human versus AI) can change under plausible changes to the relative weighting of false positive versus false negative costs. Based on our findings, we discuss implications for how to reliably evaluate the performance of human-AI teams.

2. Related Work

A substantial body of interdisciplinary work has investigated the relative performance of human versus AI decision policies (e.g., see meta-analyses [1, 19, 20]). More recently, HCI and ML researchers have examined *AI-assisted human decision-making*, studying how AI tools and process interventions (e.g., explanations) might augment human-only decisions [21, 22, 23]. Although studies suggest that AI-only decisions often outperform human-only decisions [13?], **these comparisons typically rely on imperfect proxy outcomes that may overlook the impacts of outcome measurement error.**

Recent work has identified key challenges affecting the validity of proxy labels used to measure the performance of decision policies [9, 4]. One key issue is outcome measurement error, in which proxy outcomes systematically differ from the target outcome of interest to humans [10, 24]. A second challenge is *omitted payoff bias*, in which the broader objectives of human decision-makers are imperfectly reflected by human-AI decision performance metrics [25, 4]. A specific source of omitted payoff bias is the cost ratio reflecting the utility of false positive versus false negative classification outcomes — e.g., the cost of ordering a test for a sick patient (false positive) versus turning away a healthy patient (false negative). While many evaluations weigh these costs equally (e.g., via accuracy), it is critical to examine different cost ratios that may encode differing objective functions. In this work, we develop an approach for assessing the impact of both of these factors on policy comparisons.¹

3. Framework Overview

We now introduce key components of our framework here and defer full discussion to Appendix A. Our framework examines the sensitivity of decision policy comparisons to measurement error via a **measurement sensitivity analysis** [24]. This method identifies the largest magnitude of measurement error permissible before the relative performance of two policies is no-longer definitive. In the case of our framework, we measure the difference between two decision policies π_1 and π_2 via the regret $R(\pi_1, \pi_2) = V(\pi_1) - V(\pi_2)$, where $V(\pi_1)$ is a measure of the performance of each independent policy.

Let Y^* denote a binary unobserved target outcome of interest and let Y denote an observed proxy outcome. The sensitivity parameter $\gamma = \mathbb{P}(Y^* \neq Y)$ controls the probability that the target and proxy outcomes differ. This sensitivity parameter can be used to construct a performance interval

¹See Section A for a comprehensive discussion of related work.

Table 1

An overview of policy comparisons included in our empirical analysis.

Prior Work	Domain	Proxy Outcomes	First policy (π_1)	Second policy (π_2)	Reported findings
Dressel and Farid [21] Lin et al. [22] Biswas et al. [15] Fogliato et al. [13]	Judicial	Y_1 : General re-arrest Y_2 : Violent re-arrest	Human: MTurk recidivism predictions	Algorithm: Thresholded risk scores	Algorithm better than Human
Green and Chen [23] Kawakami et al. [31]	Judicial / Lending Child welfare	Y_1 : General re-arrest / Y_1 : Loan default Y_1 : Placement, Y_2 : Re-referral, Y_3 : Services	Human+Algorithm: Participant decisions after viewing risk scores	Algorithm: Thresholded risk scores	Algorithm better than Human+Algorithm
Obermeyer et al. [18] Moody et al. [25]	Healthcare	Y_1 : Diagnosis with new chronic condition Y_1 : Hospitalization	Human+Algorithm: Physician decisions after reviewing risk scores Scoring Rule: Modified Early Warning Score (≥ 5)	Algorithm: Thresholded risk scores	Algorithm worse than Human+Algorithm Algorithm better than Scoring Rule

$\underline{R}(\pi_1, \pi_2; \gamma)$, $\bar{R}(\pi_1, \pi_2; \gamma)$ that contains the best-case and worst-case performance difference between π_1 and π_2 for a given magnitude of measurement error (γ). The measurement sensitivity analysis identifies the smallest magnitude of γ permissible before the performance interval contains zero.

Our second methodology, **cost-ratio decision curves** examine the relative performance of decision policies for different misclassification costs. Let c_0 denote the cost of a false positive and let c_1 denote the cost of a false negative. The cost ratio $\eta = \frac{c_0}{c_1}$ denotes the relative cost of the two costs. Cost ratio decision-curves plot the regret interval $\underline{R}(\pi_1, \pi_2; \gamma, \eta)$, $\bar{R}(\pi_1, \pi_2; \gamma, \eta)$ for different values of η .

4. Re-examining the Performance of Hybrid Human-AI Teams

We now apply our framework to revisit decision policy comparisons studied in human-algorithm decision-making literature. Our analysis illustrates how sensitivity analyses and decision curves can be used to examine the robustness of performance comparisons to imperfect proxy outcomes. Critically, our results indicate that (1) **policy comparisons may be inconclusive under a small and plausible magnitude of measurement error** and (2) the optimal decision policy **can change under plausible shifts to the relative weighting of classification costs**. Table 1 summarizes policy comparisons included in our empirical application. We analyze data from prior human-subjects experimental studies in judicial [13, 14, 24, 16], lending [16], and child welfare domains [17]. We provide an overview of the judicial and healthcare domains below and detail additional domains in Appendix B.

Judicial: In the criminal justice context, AI models have been introduced to inform judicial pre-trial release decisions [26, 27]. While judges are often interested in the risk of *re-offense* (Y^*) if a defendant is released, we instead observe *re-arrest* outcomes in administrative data. Re-arrest is an imperfect measure of re-offense because many crimes go un-reported or un-addressed by police (see Section A.3.1). Re-arrest outcomes are sometimes segmented into general (i.e., all charge types) and violent (e.g., Murder). We use these outcomes for our decision curves in Section A.4.²

²While we do not necessarily endorse the use of AI models in this domain (or others included below), human versus AI

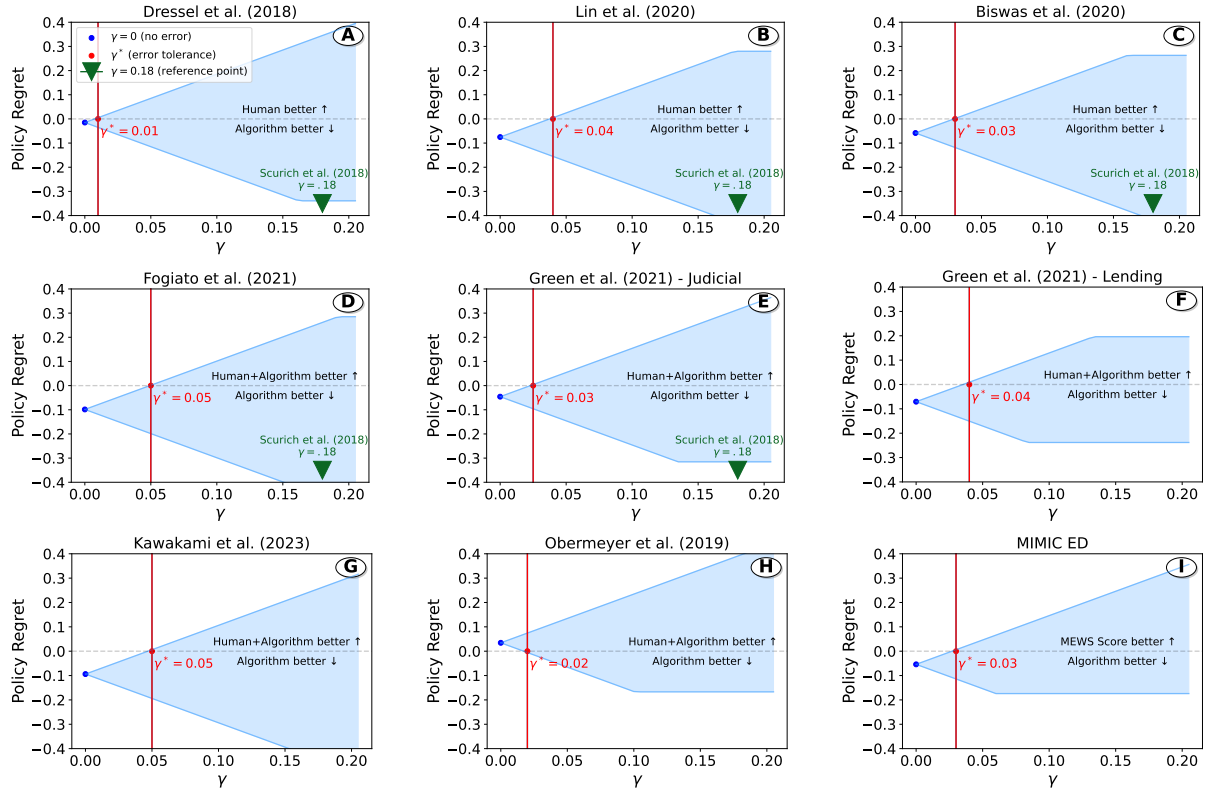


Figure 1: Measurement sensitivity analyses. Blue point ($\gamma = 0$) indicates observed performance on proxy labels under no measurement error. Sensitivity parameter γ^* in red indicates the largest magnitude of measurement error permissible before the comparison is inconclusive. Green marker shows a reference point for a plausible measurement error tolerance in criminal justice contexts based on the findings of Scurich and John [28]. The shape of regret bounds is non-linear in γ for plots E, F, H and I due to the constraint that γ does not exceed the size of the high-cost probability region (see e.q. 1).

Healthcare: AI models are increasingly used in clinical contexts to help doctors decide which patients should be enrolled in preventative care programs. However, observed labels such as cost of medical care (Y) are often an imperfect reflection of medical need (Y^*). We re-analyze a policy comparison provided by Obermeyer et al. [11] by comparing human+AI versus AI-only decisions.

4.1. Results: Measurement Sensitivity Analyses

Figure 1 shows measurement sensitivity analyses for policy comparisons reported in Table 1. The y-axis titled policy regret refers to the difference in cost of an AI-only versus alternative policy (e.g., Human, Human+AI). A negative policy regret indicates that the AI-only policy performs best. The blue point ($\gamma = 0$) indicates the observed performance difference without measurement error, while γ^* indicates the largest magnitude of measurement error allowed before the comparison is inconclusive. We provide a plausible reference point for measurement error in the criminal justice context [28].³

Critically, we find that all studies have a small error tolerance. Panel D of Figure 1 has the *largest* measurement error tolerance among all studies in the criminal justice domain, with $\gamma^* = 0.05$. **This plot indicates that the relative performance of Human versus AI policies compared by Fogliato et al. [15] is inconclusive if over 5% of outcomes are mismeasured.** This tolerance is well-within the plausible range of prior studies on measurement error in re-arrest outcomes (Section A.3.2). For example, Scurich and John [28] estimate that *at least 18% of re-offenses are incorrectly recorded in official*

comparisons in criminal justice (e.g., [13? , 14, 15]) have shaped the conversation surrounding the relative benefits of human versus AI decisions. This makes it especially critical to re-examine the sensitivity of these studies to imperfect proxy labels.

³Figure 1 isolates the impacts of measurement error by fixing $\eta = 1$ and varying γ .

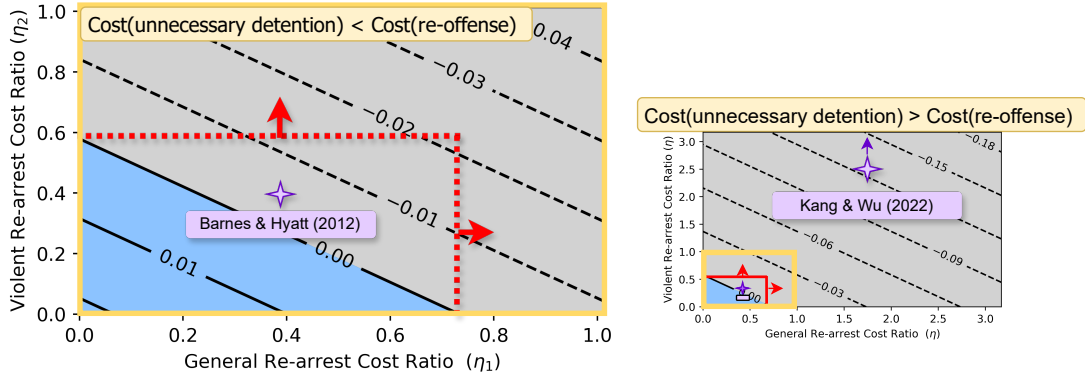


Figure 2: Cost-ratio decision curve for the study reported by Dressel and Farid [13]. The larger left panel is an expanded version of the highlighted section in the right panel. The AI-only policy is better than the human-only policy when the cost of unnecessary detentions (i.e., false positives) is greater than the cost of re-offense (i.e., false negatives). When the cost of unnecessary detentions is less than the cost of re-offense, the optimal policy depends upon the specific ratio across both violent re-arrest and general re-arrest outcomes.

arrest records, which exceeds the 5% tolerance reported in Panel D of Figure 1 by a large margin. While it is critical to link measurement error estimates to the specific sample being used for a policy comparison (see Section A.3.1), these results indicate that prior human-subjects experimental studies comparing human versus AI policies in criminal justice may be inconclusive under outcome measurement error. A similar analysis can be performed to assess whether the small measurement tolerance reported on other domains (e.g., education, healthcare) falls within a plausible range.

4.2. Results: Cost Ratio Decision Curves

Figure 4 shows a decision curve for the study conducted by Dressel and Farid [13]. The bottom right panel shows policy regret over regions where the cost of unnecessary detentions (i.e., *false positives*) exceeds the cost of defendant re-offense (i.e., *false negatives*). Negative policy regret over this region indicates that the AI-only policy performs better than the human-only policy under this relationship for classification costs. Kang and Wu [29] found that members of the public preferred a general re-arrest cost ratio of 1.67 false positives to one false negative ($\eta = 1.67$) and a violent re-arrest cost ratio of seven false positives to one false negative ($\eta = 7$), which falls within this region.

The upper left panel of Figure 2 illustrates the setting in which the cost of unnecessary detentions is less than the cost of re-offenses. This region shows that the AI-only policy performs better than the Human-only policy if judges are willing to accept .6+ unnecessary detentions for each *violent* re-arrest prevented OR we were willing to accept .7+ unnecessary detentions for every *general* re-arrest prevented. The AI-only policy also out-performs the Human-only policy for combinations of classification costs spanning ($\eta_1 = 0, \eta_2 = 1$) to ($\eta_1 = .75, \eta_2 = 0$). The classification cost ratio which Barnes and Hyatt [30] elicit from criminal justice stakeholders falls within this region of the multi-outcome decision curve ($\eta = 0.4$). Taken together, this evidence indicates that the most likely set of classification cost preferences falls within the grey region of Figure 2 – i.e., the AI-only policy likely out-performs alternative policies in this setting. Appendix B provides additional cost ratio analysis for studies listed in Table 1.

Our analysis is not intended to provide a definitive recommendation for a feasible set of classification costs in this domain. Instead, classification cost ratios are an important public policy decision which should be made by domain experts and impacted community members. However, our analysis underscores that these tacit decisions imply qualitatively different conclusions when performing comparative performance analyses. **This underscores the need to characterize sources of outcome variable uncertainty while comparing hybrid human-AI team configurations, in addition to communication, coordination, and collaboration interventions designed to improve performance.**

5. Discussion

In this work, we propose an evaluation framework for comparing decision policies under imperfect proxy labels and applied it to a diverse set of policy comparisons performed in prior literature. Our analysis shows that (1) a small magnitude of outcome measurement error can yield inconclusive policy performance comparisons, and (2) the qualitative conclusions of a performance comparison can depend on the relative weighting of misclassification costs. We conclude by discussing implications for analyses of lab-based empirical studies before highlighting limitations of our framework.

Our empirical application cites several experimental human subjects studies assessing the relative performance of human versus algorithmic decision-making. While the directionality of our results is consistent with that of prior studies (i.e., algorithmic policies tend to out-perform alternatives [19, 1, 20]), our results indicate that human versus algorithmic performance comparisons may be less conclusive than previously imagined. In particular, because algorithms are conducted on *imperfect proxy labels*, standard performance analyses (e.g., AU-ROC and Accuracy) will naturally tend to favor the algorithm. Yet, this performance assessment overlooks sources of uncertainty that impact outcome variables. **Going forward, it may be necessary to carefully re-examine prior comparisons of human versus algorithmic decision-making to assess their robustness to measurement error.**

Our formulation simplifies assumptions to cleanly parameterize target-variable related uncertainty sources that impact policy performance comparisons. Most notably, we assume ignorability (i.e., no unmeasured confounding), which can introduce additional uncertainty in off-policy evaluation of decision policies. We also assume that outcomes are observed under both possible actions being considered by π_1 and π_2 . While this simplification overlooks known challenges such as selective labels [31], our analysis foregrounds uncertainty arising from under-specified classification costs and outcome measurement error because (1) policy evaluation under confounding and unobserved counterfactuals has been studied by prior work [3, 32] and (2) would introduce additional uncertainty sources which could yield vacuous regret intervals. As a result, our results should be interpreted as a *lower bound* on uncertainty impacting policy performance comparisons.

Furthermore, our partial identification result recovers conservative worst-case regret bounds. An advantage this bound is that it subsumes more specific measurement error models (e.g., group-dependent error [33]) which may arise in particular policy evaluation contexts. However, our bounds may also be overly conservative in some contexts because they do not exploit additional domain specific information which could be used to tighten regret intervals. For instance, assuming that measurement errors differentially impact the recommendations of the algorithmic policy would yield tighter regret intervals and enable certifying policy performance differences up to a larger magnitude of error.

6. Declaration on Generative AI

A Generative AI system (ChatGPT) was used for (i) sentence polishing and (ii) rephrasing of some sentences. Specifically, existing content written by a human was provided to the model with a prompt instructing the model to identify opportunities to improve clarity, conciseness, style, and identify spelling mistakes. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] W. M. Grove, D. H. Zald, B. S. Lebow, B. E. Snitz, C. Nelson, Clinical versus mechanical prediction: a meta-analysis., *Psychological assessment* 12 (2000) 19.
- [2] R. S. Baker, A. Hawn, Algorithmic bias in education, *International Journal of Artificial Intelligence in Education* (2021) 1–41.
- [3] A. Rambachan, et al., Identifying prediction mistakes in observational data, *Harvard University* (2021).

- [4] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions, *The quarterly journal of economics* 133 (2018) 237–293.
- [5] S. Mullainathan, Z. Obermeyer, A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions, *National Bureau of Economic Research*, 2019.
- [6] P. Spitzer, J. Holstein, P. Hemmer, M. Vössing, N. Kühl, D. Martin, G. Satzger, On the effect of contextual information on human delegation behavior in human-ai collaboration, *arXiv preprint arXiv:2401.04729* (2024).
- [7] P. Hemmer, M. Schemmer, N. Kühl, M. Vössing, G. Satzger, Complementarity in human-ai collaboration: Concept, sources, and evidence, *arXiv preprint arXiv:2404.00029* (2024).
- [8] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–16.
- [9] L. Guerdan, A. Coston, Z. S. Wu, K. Holstein, Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 688–704.
- [10] R. Fogliato, A. Xiang, Z. Lipton, D. Nagin, A. Chouldechova, On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes, in: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 100–111.
- [11] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* 366 (2019) 447–453.
- [12] A. L. Coston, A. Kawakami, H. Zhu, K. Holstein, H. Heidari, A validity perspective on evaluating the justified use of data-driven decision-making algorithms, *First IEEE Conference on Secure and Trustworthy Machine Learning* (2022).
- [13] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, *Science advances* 4 (2018) eaao5580.
- [14] A. Biswas, M. Kolczynska, S. Rantanen, P. Rozenshtein, The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions, in: *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 97–104.
- [15] R. Fogliato, A. Chouldechova, Z. Lipton, The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies, *Proceedings of the ACM on Human-Computer Interaction* 5 (2021) 1–24.
- [16] B. Green, Y. Chen, Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts, *Proceedings of the ACM on Human-Computer Interaction* 5 (2021) 1–33.
- [17] A. Kawakami, L. Guerdan, Y. Cheng, K. Glazko, M. Lee, S. Carter, N. Arechiga, H. Zhu, K. Holstein, Training towards critical use: Learning to situate ai predictions relative to human knowledge, in: *Proceedings of The ACM Collective Intelligence Conference*, 2023, pp. 63–78.
- [18] G. B. Moody, R. G. Mark, A. L. Goldberger, Physionet: a web-based resource for the study of physiologic signals, *IEEE Engineering in Medicine and Biology Magazine* 20 (2001) 70–75.
- [19] P. E. Meehl, Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. (1954).
- [20] S. Ægisdóttir, M. J. White, P. M. Spengler, A. S. Maugherman, L. A. Anderson, R. S. Cook, C. N. Nichols, G. K. Lampropoulos, B. S. Walker, G. Cohen, et al., The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction, *The Counseling Psychologist* 34 (2006) 341–382.
- [21] V. Lai, C. Tan, On human predictions with explanations and predictions of machine learning models: A case study on deception detection, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 29–38.
- [22] B. Green, Y. Chen, The principles and limits of algorithm-in-the-loop decision making, *Proceedings of the ACM on Human-Computer Interaction* 3 (2019) 1–24.
- [23] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, *Proceedings of the ACM on Human-Computer*

Interaction 5 (2021) 1–21.

- [24] R. Fogliato, A. Chouldechova, M. G'Sell, Fairness evaluation in presence of biased noisy labels, in: International conference on artificial intelligence and statistics, PMLR, 2020, pp. 2325–2336.
- [25] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, S. Mullainathan, Productivity and selection of human capital with machine learning, *American Economic Review* 106 (2016) 124–127.
- [26] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, *Northpointe Inc* 7 (2016).
- [27] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, S. Venkatasubramanian, It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks, *arXiv preprint arXiv:2106.05498* (2021).
- [28] N. Scurich, R. S. John, The dark figure of sexual recidivism, *Behavioral Sciences & the Law* 37 (2019) 158–175.
- [29] B. Kang, S. Wu, False positives vs. false negatives: public opinion on the cost ratio in criminal justice risk assessment, *Journal of Experimental Criminology* 19 (2023) 919–941.
- [30] G. C. Barnes, J. M. Hyatt, Classifying adult probationers by forecasting future offending, *National Institute of Justice*. Retrieved February 4 (2012) 2020.
- [31] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, S. Mullainathan, The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 275–284.
- [32] A. Rambachan, A. Coston, E. Kennedy, Counterfactual risk assessments under unmeasured confounding, *arXiv preprint arXiv:2212.09844* (2022).
- [33] J. Wang, Y. Liu, C. Levy, Fair classification with group-dependent label noise, in: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 526–536.
- [34] K. Z. Gajos, L. Mamykina, Do people engage cognitively with ai? impact of ai assistance on incidental learning, in: *27th International Conference on Intelligent User Interfaces*, 2022, pp. 794–806.
- [35] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, M. Terry, Onboarding materials as cross-functional boundary objects for developing ai assistants, in: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.
- [36] V. Lai, H. Liu, C. Tan, "why is' chicago'deceptive?" towards building model-driven tutorials for humans, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.
- [37] B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm aversion: people erroneously avoid algorithms after seeing them err., *Journal of Experimental Psychology: General* 144 (2015) 114.
- [38] L. Cheng, A. Chouldechova, Heterogeneity in algorithm-assisted decision-making: A case study in child abuse hotline screening, *Proceedings of the ACM on Human-Computer Interaction* 6 (2022) 1–33.
- [39] H.-F. Cheng, L. Stapleton, A. Kawakami, V. Sivaraman, Y. Cheng, D. Qing, A. Perer, K. Holstein, Z. S. Wu, H. Zhu, How child welfare workers reduce racial disparities in algorithmic decisions, in: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (forthcoming), 2022.
- [40] M. De-Arteaga, R. Fogliato, A. Chouldechova, A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [41] A. Chouldechova, D. Benavides-Prado, O. Fialko, R. Vaithianathan, A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, in: *Conference on Fairness, Accountability and Transparency*, PMLR, 2018, pp. 134–148.
- [42] R. M. Dawes, The robust beauty of improper linear models in decision making, in: *Rationality and Social Responsibility*, Psychology Press, 2008, pp. 321–344.
- [43] L. Guerdan, A. Coston, K. Holstein, Z. S. Wu, Counterfactual prediction under outcome measurement error, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and*

Transparency, 2023, pp. 1584–1598.

- [44] B. Butcher, C. Robinson, M. Zilka, R. Fogliato, C. Ashurst, A. Weller, Racial disparities in the enforcement of marijuana violations in the us, in: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 130–143.
- [45] A. Kawakami, V. Sivaraman, H.-F. Cheng, L. Stapleton, Y. Cheng, D. Qing, A. Perer, Z. S. Wu, H. Zhu, K. Holstein, Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support, in: *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.
- [46] S. Mullainathan, Z. Obermeyer, Does machine learning automate moral hazard and error?, *American Economic Review* 107 (2017) 476–480.
- [47] J. Watson-Daniels, S. Barocas, J. M. Hofman, A. Chouldechova, Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 297–311.
- [48] M. Zanger-Tishler, J. Nyarko, S. Goel, Risk scores, label bias, and everything but the kitchen sink, *arXiv preprint arXiv:2305.12638* (2023).
- [49] R. Fogliato, A. K. Kuchibhotla, Z. Lipton, D. Nagin, A. Xiang, A. Chouldechova, Estimating the likelihood of arrest from police records in presence of unreported crimes, *arXiv preprint arXiv:2310.07935* (2023).
- [50] S. Mullainathan, Z. Obermeyer, On the inequity of predicting a while hoping for b, in: *AEA Papers and Proceedings*, volume 111, 2021, pp. 37–42.
- [51] M. De-Arteaga, A. Dubrawski, A. Chouldechova, Leveraging expert consistency to improve algorithmic decision support, *arXiv preprint arXiv:2101.09648* (2021).
- [52] A. Coston, A. Mishler, E. H. Kennedy, A. Chouldechova, Counterfactual risk assessments, evaluation, and fairness, in: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 582–593.
- [53] A. J. Vickers, E. B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, *Medical Decision Making* 26 (2006) 565–574.
- [54] M. Fitzgerald, B. R. Saville, R. J. Lewis, Decision curve analysis, *Jama* 313 (2015) 409–410.
- [55] M. Feffer, M. Skirpan, Z. Lipton, H. Heidari, From preference elicitation to participatory ml: A critical survey & guidelines for future research, in: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 38–48.
- [56] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 375–385.
- [57] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika* 70 (1983) 41–55.
- [58] P. R. Rosenbaum, Sensitivity analysis for certain permutation inferences in matched observational studies, *Biometrika* 74 (1987) 13–26.
- [59] P. R. Rosenbaum, Sensitivity analysis in observational studies, *Encyclopedia of statistics in behavioral science* (2005).
- [60] Z. Tan, A distributional approach for causal inference using propensity scores, *Journal of the American Statistical Association* 101 (2006) 1619–1637.
- [61] I. Díaz, M. J. van der Laan, Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems, *The international journal of biostatistics* 9 (2013) 149–160.
- [62] E. Ben-Michael, K. Imai, Z. Jiang, Policy learning with asymmetric counterfactual utilities, *Journal of the American Statistical Association* (2023) 1–25.
- [63] E. Ben-Michael, D. J. Greiner, K. Imai, Z. Jiang, Safe policy learning through extrapolation: Application to pre-trial risk assessment, *arXiv preprint arXiv:2109.11679* (2021).
- [64] K. Hardin, N. Scurich, The dark figure of violence committed by discharged psychiatric inpatients, *The Journal of Forensic Practice* 24 (2022) 229–240.
- [65] R. Fogliato, A. K. Kuchibhotla, Z. Lipton, D. Nagin, A. Xiang, A. Chouldechova, Estimating the likelihood of arrest from police records in presence of unreported crimes, *The Annals of Applied Statistics* 18 (2024) 1253–1274.

- [66] S. L. Hui, S. D. Walter, Estimating the error rates of diagnostic tests, *Biometrics* (1980) 167–171.
- [67] L. H. Aiken, S. P. Clarke, D. M. Sloane, I. H. O. R. Consortium, Hospital staffing, organization, and quality of care: cross-national findings, *International Journal for quality in Health care* 14 (2002) 5–14.
- [68] A. Coşer, M. M. Maer-matei, C. Albu, Predictive models for loan default risk assessment., *Economic Computation & Economic Cybernetics Studies & Research* 53 (2019).
- [69] J. N. Day, Credit, capital and community: informal banking in immigrant communities in the united states, 1880–1924, *Financial History Review* 9 (2002) 65–78.
- [70] D. Low, What triggers mortgage default? new evidence from linked administrative and survey data, *Review of Economics and Statistics* (2023) 1–26.
- [71] R. Vaithianathan, E. Putnam-Hornstein, N. Jiang, P. Nand, T. Maloney, Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation, *Center for Social data Analytics* (2017).
- [72] H.-F. Cheng, L. Stapleton, A. Kawakami, V. Sivaraman, Y. Cheng, D. Qing, A. Perer, K. Holstein, Z. S. Wu, H. Zhu, How child welfare workers reduce racial disparities in algorithmic decisions, in: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–22.
- [73] C. P. Subbe, M. Kruger, P. Rutherford, L. Gemmel, Validation of a modified early warning score in medical admissions, *Qjm* 94 (2001) 521–526.
- [74] C. Elkan, The foundations of cost-sensitive learning, in: *International joint conference on artificial intelligence*, volume 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

A. Extended Related Work

In this section, we introduce a body of literature which performance comparative performance assessments of decision policies (Section A.1) which frequently overlooks key sources of uncertainty impacting proxy outcomes (Section A.2). We then introduce several model evaluation approaches (i.e., decision curves, sensitivity analyses) which we extend to support better-informed policy comparisons under imperfect proxy outcomes (Section A.3).

A.1. Comparisons of human versus algorithmic decision-making policies

A large body of research has investigated the relative performance of human versus algorithmic decision-making approaches [1, 19, 20, 13?, 21, 22, 16, 23, 34, 35, 36, 15, 37]. This work dates back to influential meta-analyses of studies comparing the predictive performance of “*clinical*” (i.e., human expert) versus “*actuarial*” (i.e., statistical model) decisions across a variety of decision-making domains [1, 19, 20]. ProPublica’s widely-cited article on the COMPAS risk assessment spurred renewed interest in the relative accuracy and fairness of human versus algorithmic decision-making in the criminal justice context [13?].

More recently, a growing body literature studying algorithm-assisted *human* decision-making investigates how algorithmic decision support tools might augment existing (i.e., human only) protocols [21, 22, 16, 23, 34, 35, 36, 15, 37, 17]. Some studies examine the quality of decisions made in real-world system deployments via retrospective analyses of log data [38, 39, 40, 41], while others perform experimental human subjects studies examining how various interventions (e.g., training, explanations, workflow modifications) impact decision performance under controlled conditions [22, 16, 23, 34, 35, 36, 15, 37]. While many studies compare decision policies in aggregate (i.e., averaged over multiple human decision-makers), some report more granular breakdowns at the individual decision-maker level [38, 3, 42].

To date, the prevailing narrative in this literature is that algorithmic approaches tend to make higher-quality decisions than humans would acting alone [1, 19, 20, 13?]. **However, prior studies typically assess the quality of decisions via predictive performance computed via imperfect proxy outcomes. As a result, existing policy comparisons overlook key sources of uncertainty impacting target variables in real-world predictive modeling contexts.**

A.2. Sources of uncertainty impacting proxy outcomes

Recent research has identified a host of challenges impacting the validity of target variables in real-world predictive modeling contexts [9, 4, 43, 27, 44, 45, 39, 11, 46, 5, 25, 10, 24, 47, 48, 33]. For example, predictive models deployed to inform judicial pre-trial release decisions often target defendant re-arrest as a proxy for re-offense [10, 49]. Predictive models intended to inform healthcare program enrollment decisions often target measures of utilization (e.g., cost) as a proxy for medical need [11, 50]. *Outcome measurement error* can impact performance evaluations when readily-available proxy outcomes (e.g., cost, re-arrest) systematically differ from the target outcome of policy interest (e.g., health need, re-offense) [9]. Similarly, *omitted payoff bias* occurs when the objectives of human decision-makers (e.g., the relative weighting of misclassification outcomes) are incompletely captured by decision policy performance measures [25, 4, 51, 16, 3]. While additional challenges (e.g., intervention effects [52], unmeasured confounding [32]) can also impact policy performance comparisons, the framework we develop in this work assesses impacts of challenges most directly-related to proxy labels (i.e., outcome measurement error, omitted payoffs) on policy performance comparisons.

A.3. Policy performance evaluation under uncertainty

In this work, we build upon existing evaluation tools (i.e., decision curves [53, 54], sensitivity analysis [24]) to support better-informed policy performance comparisons under imperfect proxy outcomes.

A.3.1. Assessing omitted payoff bias

A large body of *participatory machine learning* literature leverages feedback from non-technical stakeholders to refine development and evaluation of AI models [55]. In our context, utility elicitation methods have been developed to (1) assess stakeholder perceptions surrounding the relative cost of false positive and false negative classification outcomes and (2) leverage these perceptions to select appropriate risk thresholds [29, 30]. For example, Kang and Wu [29] explore public perceptions of the misclassification cost ratio by performing an online experimental study which asks participants to rank models with varying false positive and false negative rates. Their analysis found that participants preferred a cost ratio of 1.67 false positives (i.e., unnecessary detentions) for each false negative for general (i.e., all purpose) recidivism outcomes and 7 false positives for each false negative for violent recidivism outcomes.

Decision curve analysis is an evaluation approach frequently used to compare predictive models against relevant baselines (e.g., test all patients, test no patients) across a *range* of decision thresholds in medical contexts [53, 54]. Rambachan et al. [3] leverages a variant of this technique in the criminal justice domain to compare judicial decisions against algorithmic decisions over a variety of classification costs. In this work, we extend this approach to (1) compare performance over *multiple policy-relevant target outcomes* and (2) model interactions with outcome measurement error impacting target variables.

A.3.2. Assessing outcome measurement error

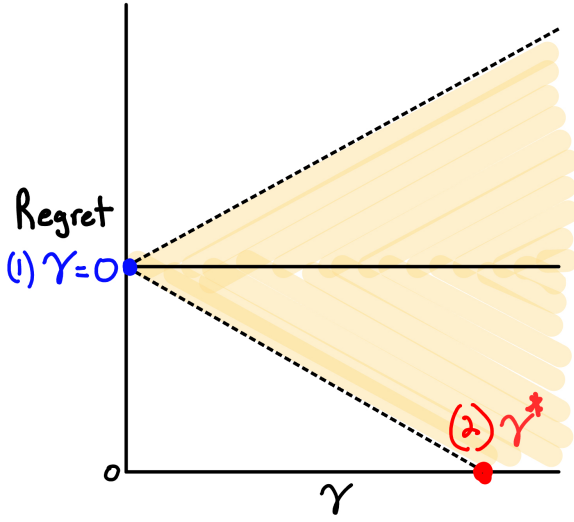
In many contexts, it is difficult to characterize the precise relationship between the unobserved target outcome of interest (e.g., crime, health) and the proxy observed in performance benchmarking datasets (e.g., re-arrest, cost of care) [56]. This is because linking unobserved constructs with observed proxies requires measurement assumptions which cannot be directly verified. However, sensitivity analyses offer a practical alternative to model evaluation under measurement error without need for strong assumptions on the relationship between target and proxy outcomes.

Sensitivity analyses assess the magnitude of bias in analysis required to draw the overall findings (e.g., the relative performance of decision policies) into question. This analysis has been widely-leveraged in causal inference settings to assess the robustness of causal effects to confounding [57, 58, 59, 60, 32] or measurement error [61]. Most directly-related to our setting, Fogliato et al. [24] develop a framework for characterizing the predictive performance of risk assessments under outcome measurement error. We extend this approach to characterize the *relative performance of two decision policies* rather than a single policy in isolation. Additionally, while [24] study traditional predictive performance measures (e.g., AU-ROC, FPR) we study a cost-sensitive performance measure which assesses the differential-impact of false positive and false negative classification errors on the comparative performance analysis.

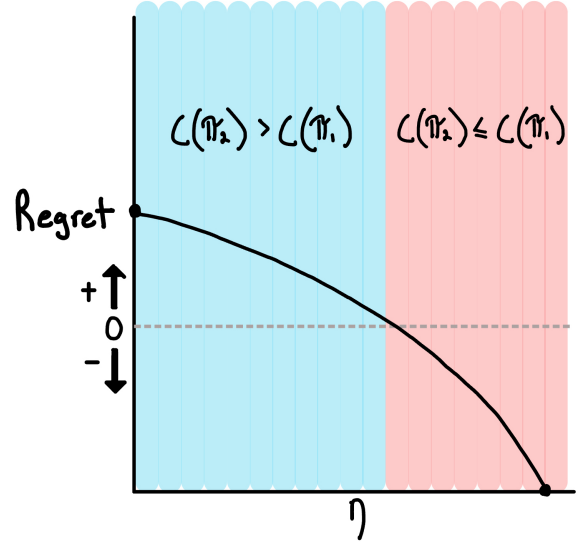
A. Methodological Framework

In this appendix, we introduce our approach for comparing decision policies under imperfect proxy labels. This framework characterizes the impact of two distinct sources of uncertainty – outcome measurement error and under-specified classification costs – on policy performance comparisons. We introduce our technical approach in Section A.2. We then introduce two evaluation tools leveraging this approach:

- **Measurement sensitivity analysis** which characterizes the robustness of policy comparisons to outcome measurement error (Section A.3).
- **Cost ratio decision curves** which assess the robustness of policy comparisons to the relative cost of false positive and false negative classification errors (Section A.4)



(a) Illustration of a measurement sensitivity analysis plotting upper and lower regret bounds as a function of the measurement error tolerance. Yellow region corresponds to feasible values of the true regret under a γ -magnitude of measurement error. The far left point shown in blue ($\gamma = 0$) corresponds to the regret computed over proxy labels. The far right point shown in red (γ^*) shows the maximum amount of measurement error permissible before the regret interval contains zero. This plot assumes that \mathbf{c} is fixed and we vary γ over a single proxy.



(b) Illustration of a cost-ratio decision curve plotting regret as a function of false positive to false negative cost ratio. Blue region indicates cost ratios $\eta = \frac{c_{1,0}}{c_{0,1}}$ for which π_2 has a higher cost than π_1 . Red region indicates cost ratios for which π_2 has lower cost than π_1 . The rank-order of model performance reverses when the regret crosses zero. This plot assumes that the cost of true positives and true negatives are fixed.

A.1. Preliminaries

Let $\pi_1 : \mathcal{X} \rightarrow \mathcal{A}$ and $\pi_2 : \mathcal{X} \rightarrow \mathcal{A}$ be two decision policies which assign binary actions $\mathcal{A} \in \{0, 1\}$ given covariates $X \in \mathcal{X}$. Let D_1 and D_2 denote random variables indicating actions selected under π_1 and π_2 , respectively.⁴ Let $Y^* \in \{0, 1\}^K$ be a vector of K *unobserved* target outcomes and let $Y \in \{0, 1\}^K$ be a vector of K corresponding *observed* proxy outcomes. We let $\mathbb{P}^*(X, D_1, D_2, Y^*, Y)$ denote the unobserved joint distribution over target and proxy outcomes and let $\mathbb{P}(X, D_1, D_2, Y)$ denote the observed joint distribution over proxy outcomes.

We quantify the performance of a policy via its *expected cost* over target outcomes. In particular, let $\mathbf{c} \in \mathbb{R}^{K \times 2 \times 2}$ be a classification cost matrix, where each entry $c_{k,a,y} \geq 0$ denotes the cost of the k 'th proxy

⁴We sometimes write π and D when defining quantities for a generic decision policy

outcome y given action a . Following the standard cost-sensitive classification setup, we assume that costs are fixed across levels of co-variates (e.g., age, race or sex).

Definition A.1 (Policy Cost). *The expected cost of π is given by*

$$C^*(\pi; \mathbf{c}) = \mathbb{E}_{\mathbf{P}^*}[c_{k,a,y} \cdot \mathbb{I}\{D = a, Y_k^* = y\}].$$

This performance measure reflects the net cost of classification outcomes computed over target variables [62, 3, 63].⁵ We can compare the relative performance of π_1 and π_2 by evaluating the *policy regret* – i.e., the difference in their expected cost.

Definition A.2 (Policy Regret). *The policy regret between π_1 and π_2 is given by*

$$R^*(\pi_1, \pi_2; \mathbf{c}) = C^*(\pi_2; \mathbf{c}) - C^*(\pi_1; \mathbf{c}).$$

A positive regret indicates that π_1 has lower cost (i.e., better performance) than π_2 . Conversely, a negative regret indicates that π_1 has higher cost (i.e., worse performance) than π_2 .

However, because target outcomes are unobserved, the policy regret is *partially identified* within an interval $\underline{R}(\pi_1, \pi_2; \mathbf{c}), \bar{R}(\pi_1, \pi_2; \mathbf{c})$ from observed proxy outcomes. We assume that each target outcome is observed under measurement error of magnitude $\gamma_k = \mathbb{P}(Y_k \neq Y_k^*)$ such that $\gamma_k < .5, \forall k$. This places a constraint on the magnitude of measurement error but makes no assumptions on which instances are mislabeled [24]. Given a vector \mathbf{c} and a finite sample $\mathcal{O} := \{(D_1^i, D_2^i, Y_1^i, \dots, Y_K^i)\}_{i=1}^n \sim \mathbb{P}(\cdot)$, our goal is to bound the oracle regret within the interval $R^*(\pi_1, \pi_2; \mathbf{c}, \gamma) \in [\underline{R}(\pi_1, \pi_2; \mathbf{c}, \gamma), \bar{R}(\pi_1, \pi_2; \mathbf{c}, \gamma)]$.

A.2. Partial identification of policy regret

In this section, we provide an approach for *partially identifying* the true policy regret within a worst-case interval using observed proxy outcomes. To ease exposition, we introduce our approach in the single outcome setting. We generalize this approach to the multiple outcome setting in Appendix C.

Recall that our measurement error model assumes up to γ percent of outcomes have been mislabeled. Therefore, we minimize regret by shifting as many instances as possible from the high-cost to the low-cost outcome. Let $w_y(d_1, d_2) = \mathbb{P}(D_1 = d_1, D_2 = d_2, Y = y)$ be the joint probability of the policy actions and the proxy outcome. Because we are interested in performance differences (i.e., *regret*), we would like to apply mislabeled examples to instances where the two policies select different actions – i.e., $w_y(0, 1)$ and $w_y(1, 0)$.

Let $\delta_{a,y} = c_{a,y} - c_{a',y}$ be the cost difference between actions a and a' for outcome y and let $c_a^* = \mathbb{I}\{\delta_{a,1} > \delta_{a,0}\}$ be the outcome which maximizes the cost under action a . We minimize regret by shifting instances from high-cost disagreement terms $w_{c_a^*}(a, 1 - a)$ to low-cost disagreement terms $w_{1-c_a^*}(a, 1 - a)$, while ensuring the number of mislabeled instances does not exceed the size of the high-cost region. In particular, we impose the constraint on γ that

$$\gamma_0 = \min\{w_{c_0^*}(0, 1), \gamma\}, \quad \gamma_1 = \min\{w_{c_1^*}(1, 0), \gamma - \gamma_0\} \quad (1)$$

for $\gamma_0 + \gamma_1 \leq \gamma$. Proposition A.1 leverages this approach to bound the true regret as a function of the user-specified error tolerance (γ) and classification cost matrix (\mathbf{c}).

Proposition A.1. *The regret $R^*(\pi_1, \pi_2)$ is bounded within the interval*

$$\begin{aligned} \underline{R}(\pi_1, \pi_2; \mathbf{c}, \gamma) &= \sum_{a,y} \delta_{a,y} \cdot (w_y(a, a') + \underline{\Gamma}(a, y)), \\ \bar{R}(\pi_1, \pi_2; \mathbf{c}, \gamma) &= \sum_{a,y} \delta_{a,y} \cdot (w_y(a, a') + \bar{\Gamma}(a, y)), \end{aligned}$$

where $\underline{\Gamma}(a, y) = (-1)^{\mathbb{I}\{y=c_a^*\}}(a \cdot \gamma_1 + a' \cdot \gamma_0)$ and $\bar{\Gamma}(a, y)$ is similarly defined by taking $c_a^* = 1 - c_a^*$.

⁵This performance measure is equivalent to the *expected welfare* or *utility* of a decision policy when utilities $u_{k,a,y} \geq 0$ are used in place of costs.

We prove this result and extend it to the multiple outcome setting in Appendix C. Importantly, the bounds we compute in Proposition A.1 assume knowledge of the user-specified cost matrix and measurement error tolerance. We now show how to use this result for policy comparisons when the exact measurement tolerance (γ) and classification costs (\mathbf{c}) are uncertain.

A.3. Measurement sensitivity analysis

The exact magnitude of measurement error impacting proxy outcomes is often unknown. As a result, policy regret bounds computed via Proposition A.1 can be overly conservative if γ is too large, or invalid if γ is too small. Fortunately, in comparative performance evaluations, we are often interested in the more tractable question “*What magnitude of measurement error is required to draw the relative performance of policies into question?*” A **measurement sensitivity analysis** operationalizes this idea by computing the smallest magnitude of error (γ^*) required to yield a regret interval containing zero (Figure ??).⁶ This analysis entails (1) varying $\gamma \in [0, .5)$, (2) computing regret bounds for each γ , then (3) computing γ^* – the smallest γ such that the regret bounds contain zero.⁷

A.3.1. Determining a “plausible” measurement error range

When interpreting a measurement sensitivity analysis, a key question involves whether γ^* falls within a “*plausible range*” of measurement error informed by domain knowledge of proxy outcomes. The results of a policy comparison conducted via proxy labels may be inconclusive if γ^* falls within a feasible range estimated by domain knowledge. A large body of scientific literature in algorithmic decision support domains (e.g., criminology, medicine) has estimated plausible ranges for outcome measurement error impacting proxy outcomes.

For example, algorithmic risk assessments deployed in criminal justice often target re-arrest as a proxy for re-offense [27, 24, 10]. In criminology, the so called “*dark figure of crime*” refers to crimes that are not recorded in official arrest statistics due to under-reporting [28]. Scurich and John [28] leverage self-reported victimization data to estimate measurement error in sexual recidivism outcomes. They find that measurement error in sexual recidivism outcomes varies between 18-41% depending on the outcome time-span and various modeling assumptions. Hardin and Scurich [64] use self-report and collateral informant data to estimate the measurement error rate violent offenses among patients discharged from psychiatric hospitals. Their analysis estimated the measurement error rate (γ) between all reporting sources (Y^*) and official records (Y) of 21%. Fogliato et al. [65] develop a methodological framework for estimating the likelihood of official police records and victimization data. Applying this framework to official records from the National Incident Based Reporting System, the authors found that on average, about 20% of violent offenses eventually result in arrest. In healthcare, similar studies are also commonly used to analyze error rates in diagnostic testing instruments [66].

A.3.2. Contextualizing measurement error estimates

When interpreting a measurement sensitivity analysis, it is critically to collect measurement error estimates that generalize to the sample used for a policy comparison. For example, a study of sexual recidivism rates (e.g., [28]) or violent re-offense rates among patients discharged from psychiatric hospitals (e.g., [64]) may not generalize to the sample of Boward County defendants used to train and evaluate the COMPAS risk assessment. Therefore, it is important to interpret measurement sensitivity parameters γ^* against a large body of contextually-similar empirical evidence.⁸

⁶In sensitivity analysis literature, γ^* is often called the *design sensitivity* or *error tolerance* because it is the cutoff value which changes the overall interpretation of the results (i.e., which policy is better overall) [59].

⁷This procedure fixes misclassification costs and varies γ . As we demonstrate in Section 4, it is also possible to vary both in combination.

⁸Some empirical studies of measurement error target the false negative rate $P(Y = 0 \mid Y^* = 1)$ and false positive rate $P(Y = 1 \mid Y^* = 0)$ of proxy outcomes while we model the sensitivity parameter $\gamma = P(Y \neq Y^*)$. Our sensitivity parameter can be recovered by multiplying by the prevalence, i.e. $\gamma = P(Y = 0 \mid Y^* = 1) \cdot P(Y^* = 1) + P(Y = 1 \mid Y^* = 0) \cdot (1 - P(Y^* = 1))$.

A.4. Cost ratio decision curves

The relative cost of classification errors (i.e., false positives, false negatives) encode additional information about a proxies' relevance to a decision-making process of interest. For example, a false negative prediction for patient mortality may be much more costly than a false positive mortality prediction, while false positive and false negative patient re-admission predictions may carry a different set of payoffs [67].

We propose a simple alternative approach to cost elicitation which characterizes the relative performance of decision policies *over a plausible range of false positive to false negative classification cost ratios* $\eta = \frac{c_{1,0}}{c_{0,1}}$. To perform this analysis, a practitioner first specifies a feasible set of false positive to false negative costs (e.g., "false negatives are at least twice as costly as false positives" or $\eta < .5$) based on domain knowledge.⁹ Next, the analyst plots the regret interval recovered by Proposition A.1 as a function of η (Figure ??). This performance analysis is *conclusive* if the direction of the performance difference is constant across all feasible values for τ . The analysis is *inconclusive* if the sign of the performance difference changes over the plausible set of cost ratios. As a result, more specific intervals for cost ratios (e.g., "a false negative is between 2 to 3 times more costly than a false positive") are more likely to yield conclusive performance comparisons.

Cost-dependent decision policies. In many settings, the decision policy being evaluated depends on the user-specified classification costs. For example, algorithmic decision policies are typically constructed by (1) estimating the conditional probability of the binary proxy outcome $\mu_k(x) = P(Y_k = 1 \mid X = x)$, then (2) thresholding predicted probabilities at a cost-dependent cut-off. Similarly, in experimental human-AI decision-making studies, it is common to construct human decision-making policies by (1) asking participants to predict the probability of a proxy outcome, then (2) assigning decisions by thresholding predicted probabilities at a cut-off [15, 16, 22].¹⁰ In the following proposition, we show how to construct an optimal decision policy for a user-specified classification cost matrix defined over multiple outcomes.

Proposition A.2. *The decision policy which minimizes cost with respect to \mathbf{c} is given by*

$$\pi^*(x) := \mathbb{I} \left\{ \sum_k \mu_k(x) \geq \tau^*(\mathbf{c}) \right\}, \text{ where}$$

$$\tau^*(\mathbf{c}) = \sum_k \frac{c_{k,1,0} - c_{k,0,0}}{c_{k,0,1} - c_{k,1,1} - c_{k,0,0} + c_{k,1,0}}$$

is the optimal probability threshold corresponding to the K proxy conditional probability functions $\mu_k(x)$.

We prove this result in Appendix C.

⁹By Proposition A.2, τ will be well-defined when $c_{0,1} = 0$. Analysts can plot regions where $c_{0,1} = 0$ in decision curves by assigning false negative costs to small values near zero ($c_{0,1} \approx e^{-10}$).

¹⁰Predicted probabilities are often bucketed into intervals of .1 or .01.

B. Additional Experimental Details and Results

Lending Domain: Financial institutions are often interested in assessing the “creditworthiness” (Y^*) of applicants for financial products [68]. In practice, organizations frequently target “loan default” (Y) as a proxy for the lending potential of an applicant. However, loan default is an imperfect measure of creditworthiness because applicants can fail to repay loans due to factors unrelated to their financial responsibility [69]. Further, the repayment timeline and type of default (e.g., technical, non-technical) can carry different implications for the lending potential of the applicant [70].

Child Welfare: Predictive models have been introduced in child welfare contexts to help social workers identify children at high-risk for abuse and neglect [41]. Often, social workers are legally required to assess imminent threats to child safety (Y^*) when deciding whether to screen-in a re-referral for further investigation. However, predictive models introduced in this context often target indirect child welfare measures (Y), such as out-of-home placement in foster care, agency re-referral, or acceptance for welfare services [45]. For example, an early version of the Allegheny Family Screening Tool targeted out-of-home placement and agency re-referral [71]. However, the re-referral outcome was later removed because it was perceived as being highly dependent on family socio-economic status and only distally related to child safety.¹¹

- Dressel and Farid [13]: We let π_1 be human risk predictions of two-year general recidivism outcomes and let π_2 be thresholded COMPAS decile scores. We analyze data from the MTurk No Race condition. We show outcomes evaluated against two-year general recidivism and violent recidivism in the two-dimensional cost ratio decision curves.
- ?]: We let π_1 be human risk predictions of two-year general recidivism outcomes and let π_2 be thresholded COMPAS decile scores. We use data from the “*Vignette, No Feedback*” condition.
- Fogliato et al. [15]: We let π_1 be human risk predictions of two-year general recidivism after viewing algorithmic risk predictions and let π_2 be algorithmic risk predictions. We analyze data from “*Phase 2, No Anchor*” condition.
- Green and Chen [16]: We let π_1 be human risk predictions of two-year general recidivism after viewing algorithmic risk predictions and let π_2 be algorithmic risk predictions. We use data from the “*Shown RA + Make Predictions*” condition in both the criminal justice and lending contexts.
- Biswas et al. [14]: We let π_1 be human risk predictions of two-year general recidivism outcomes and let π_2 be thresholded COMPAS decile scores.
- Kawakami et al. [17]: We let π_1 be human risk predictions of two-year placement in foster care and let π_2 be Allegheny Family Screening Tool scores thresholded at 15.¹² Note this data is from synthetic vignettes inspired by real-world cases. We use data from the pre-test assessment phase of the study.
- Obermeyer et al. [11]: We let π_1 be physician enrollment decisions for a high risk medical program (informed by cost of care prediction risk scores) and let π_2 be thresholded predictions of diagnosis with a new active chronic condition in the following year. Note this is a synthetic dataset matching means and covariance of the original data used by [11].
- The MIMIC-IV-ED is a database of emergency department admissions at Beth Israel Deaconess Medical Center from 2011 to 2019 [18]. In a hypothetical emergency triage task, we let the baseline policy consist of thresholded Modified Early Warning Scores (≥ 5), which is a standard checklist based scoring system used for emergency triage [73]. We compare this policy to a logistic regression model targeting hospitalization (Y). Hospitalization is an imperfect measure of patient acuity (Y^*) because some patients may be re-routed to a different facility or suffer from mortality prior to receiving care. MIMIC-IV-ED Dataset [18]: We let π_1 be the Modified Early Warning Score (MEWS) thresholded at the high-risk cutoff (≥ 5) and let π_2 be thresholded

¹¹Our analysis re-analyzes experimental human subjects data from [17], which asks participants to make decisions on *synthetic vignettes* inspired by real-world AFST cases. As a result, our analysis is not intended to be a direct assessment of the AFST, though such an assessment would be supported by our framework.

¹²Our use of 15 as a screen-in threshold follows from the high-risk protocol recommendation cutoff [72].

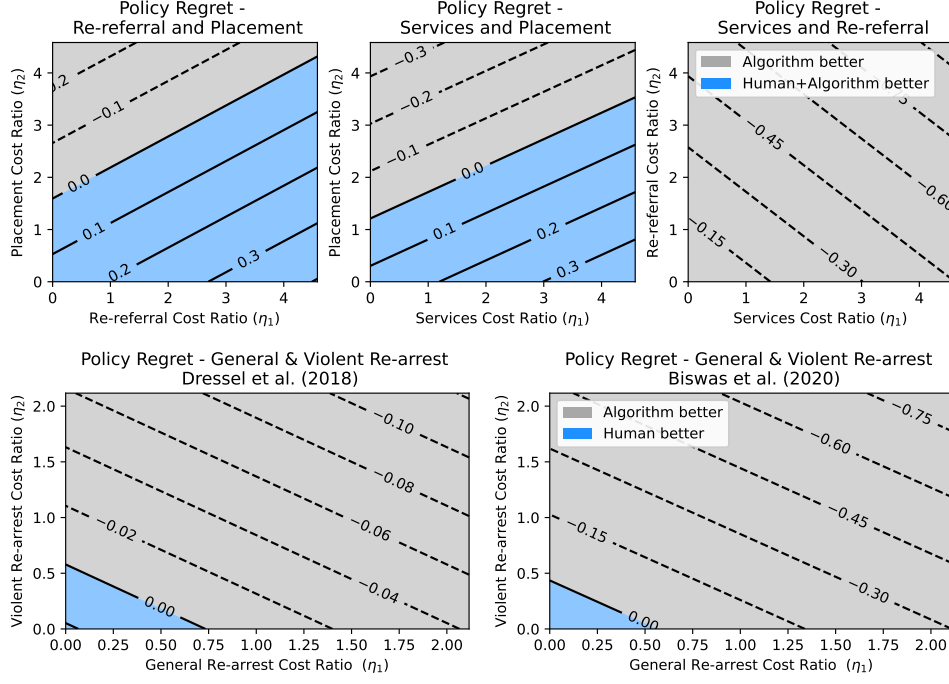


Figure 4: Multi-outcome decision curves. We fix $\gamma = 0$ and plot the policy regret as a function of the cost ratio for two pairs of two proxy outcomes. Blue indicates regions in which the Human or Human+Algorithm policy out-performs the Algorithm policy alone. Grey regions indicates regions in which the Algorithm policy performs better. Top three panels correspond to pairwise comparisons across child-welfare outcomes available as part of Kawakami et al. [17] (see main text for descriptions). Bottom two panels plot policy regret across violent and general re-arrest in the criminal justice domain.

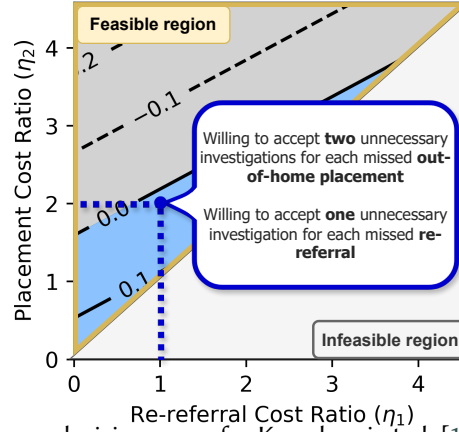


Figure 5: Analysis of multi-outcome decision curve for Kawakami et al. [17]. Infeasible region corresponds to settings in which the cost of missed out-of-home placements is less than cost of missed re-referrals. Feasible region corresponds to settings in which the cost of missed out-of-home placements is greater than cost of missed re-referrals.

hospital outcome predictions. We evaluate the scoring rule and model via hospitalization data over a held-out test set. We train a logistic regression model and evaluate both the model and the scoring rule over a held-out evaluation dataset.

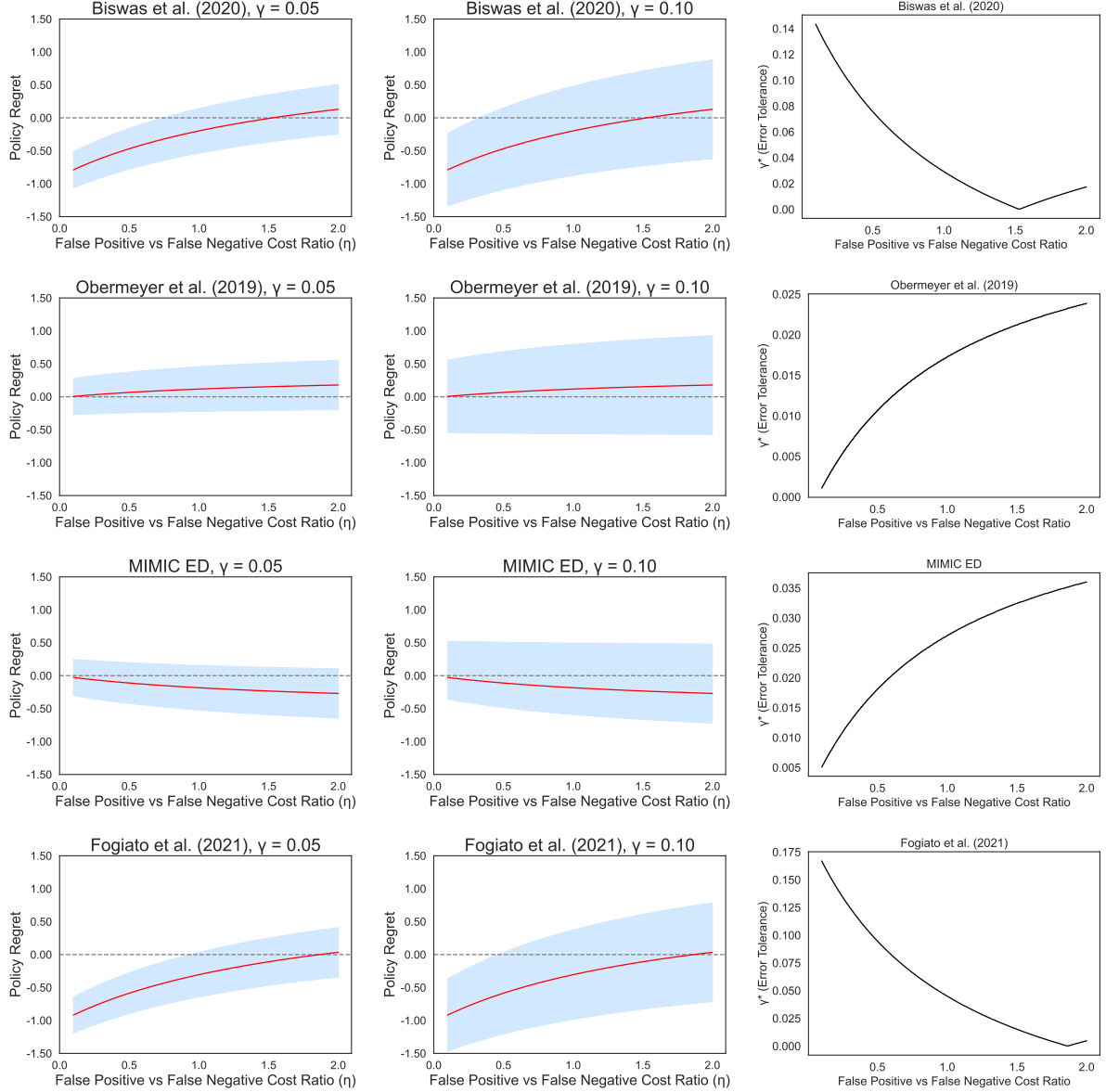


Figure 6: Measurement sensitivity decision curves for four studies included in Table ?? . Left panel fixes $\gamma = 0.05$ (left) and $\gamma = 0.15$ (center) and varies the cost ratio η . Solid black line indicates observed performance under no measurement error. Blue shaded region indicates partially-identified performance interval. Right column computes the design sensitivity parameter γ^* for each cost ratio η . We use general re-arrest outcomes for criminal justice studies [14, 15].

C. Proofs

C.1. Proposition A.1

We begin by re-defining terms in the multiple outcome setting with $K > 1$ proxies. Let $w_y(d_1, d_2) = \mathbb{P}(D_1 = d_1, D_2 = d_2, Y = y)$ be the joint probability of the policy actions and the k 'th proxy outcome. Let $\delta_{k,a,y} = c_{k,a,y} - c_{k,a',y}$ and let $c_{k,a}^* = \mathbb{I}\{\delta_{k,a,1} > \delta_{k,a,0}\}$ by the outcome with the highest cost under action a for the k 'th proxy. Additionally, let

$$\gamma_k^0 = \min\{\gamma_k, w_{c_{k,0}^*}^*(0, 1)\}, \quad \gamma_k^1 = \min\{\gamma_k - \gamma_k^0, w_{c_{k,1}^*}^*(1, 0)\}$$

be the probability mass which can be shifted from the high-cost to the low cost terms for the k 'th proxy observed under a γ_k magnitude of measurement error.

Proposition C.1. $R^*(\pi_1, \pi_2)$ is bounded within the interval

$$\begin{aligned} \underline{R}(\pi_1, \pi_2; c, \gamma) &= \sum_{k,a,y} \delta_{k,a,y} \cdot (w_{\gamma_k}(a, a') + \underline{\Gamma}(k, a, y)), \\ \bar{R}(\pi_1, \pi_2; c, \gamma) &= \sum_{k,a,y} \delta_{k,a,y} \cdot (w_{\gamma_k}(a, a') + \bar{\Gamma}(k, a, y)), \end{aligned}$$

where $\underline{\Gamma}(k, a, y) = (-1)^{\mathbb{I}\{\gamma_k = c_{k,1}^*\}}(a \cdot \gamma_k^1 + a' \cdot \gamma_k^0)$ and $\bar{\Gamma}(k, a, y)$ is similarly defined by taking $c_{k,a}^* = 1 - c_{k,a}^*$.

Proof. To begin, observe that we can factorize $C(\pi_1)$ into

$$\begin{aligned} C(\pi_1) &= \mathbb{E}_P[c_{k,a,y} \cdot \mathbb{I}\{A = a, Y_k = y\}] \\ &= \sum_{k,a,y,x} c_{k,a,y} \cdot p(A = a, Y_k = y \mid X = x) \cdot p(X = x) \\ &= \sum_{k,a,y,x} c_{k,a,y} \cdot p(A = a, Y_k = y, X = x) \\ &= \sum_{k,a,y} c_{k,a,y} \cdot p(A = a, Y_k = y) \\ &= \sum_{k,a,y} c_{k,a,y} \cdot (w_{\gamma_k}(a, 0) + w_{\gamma_k}(a, 1)). \end{aligned}$$

By the same argument, $C(\pi_2)$ factorizes into

$$C(\pi_2) = \sum_{a,y,k} c_{a,y,k} \cdot (w_{\gamma_k}(0, a) + w_{\gamma_k}(1, a)).$$

Therefore, we can express the regret over proxy outcomes as

$$\begin{aligned} R(\pi_1, \pi_2) &= C(\pi_2) - C(\pi_1) \\ &= \sum_{k,a,y} c_{k,a,y} \cdot (w_{\gamma_k}(a, a') - w_{\gamma_k}(a', a)) \\ &= \sum_{k,a,y} w_{\gamma_k}(a, a') \cdot (c_{k,a,y} - c_{k,a',y}). \end{aligned}$$

Let $c_{k,a}^* = \mathbb{I}\{\delta_{k,a,1} > \delta_{k,a,0}\}$ be the value of the proxy outcome $\gamma_k = c_{k,a}^*$ which maximizes the cost under action a . Further, let

$$\gamma_k^0 = \min\{\gamma_k, w_{c_{k,0}^*}^*(0, 1)\}, \quad \gamma_k^1 = \min\{\gamma_k - \gamma_k^0, w_{c_{k,1}^*}^*(1, 0)\}$$

be the largest probability mass which can be shifted from the high-cost terms to low-cost terms under a γ_k magnitude of error. This implies the bound

$$\begin{aligned} &\sum_{k,a} (w_{c_{k,a}^*}^*(a, a') - \gamma_{k,a}) \cdot \delta_{a,k,c_{k,a}^*} \\ &\quad + (w_{1-c_{k,a}^*}^*(a', a) + \gamma_{k,a}) \cdot \delta_{a,k,1-c_{k,a}^*} \leq R^*(\pi_1, \pi_2) \end{aligned}$$

where the result follows from re-arranging. Similarly, the upper bound follows by taking $c_{k,a}^* = 1 - c_{k,a}^*$. \square

C.2. Proposition A.2

Proof. The proof follows by extending the standard cost sensitive learning setup [74] to to our setting with K outcomes. In the single outcome case, the optimal policy is given by

$$\pi^*(x) = \arg \min \left\{ \begin{array}{l} c_{k,0,0} \cdot (1 - \mu_k(x)) + c_{k,0,1} \cdot \mu_k(x), \\ c_{k,1,0} \cdot (1 - \mu_k(x)) + c_{k,1,1} \cdot \mu_k(x) \end{array} \right\}, \quad \forall k. \quad (2)$$

The optimal probability threshold η_k^* follows by solving e.q. 2 for $\mu_k(x)$

$$\mu_k^* = \frac{c_{k,1,0} - c_{k,0,0}}{c_{k,0,1} - c_{k,1,1} - c_{k,0,0} + c_{k,1,0}}.$$

The result follows by summing over the K outcomes

$$\sum_k \mu_k^*(x) = \sum_k \frac{c_{k,1,0} - c_{k,0,0}}{c_{k,0,1} - c_{k,1,1} - c_{k,0,0} + c_{k,1,0}}$$

□