Beyond Prediction: Managing the Repercussions of Machine Learning Applications

Aline Weber*
University of Massachusetts
Amherst, MA 01003, USA
alineweber@cs.umass.edu

Blossom Metevier*
University of Massachusetts
Amherst, MA 01003, USA
bmetevier@umass.edu

Yuriy Brun
University of Massachusetts
Amherst, MA 01003, USA
brun@cs.umass.edu

Philip S. Thomas
University of Massachusetts
Amherst, MA 01003, USA
pthomas@cs.umass.edu

Bruno Castro da Silva
University of Massachusetts
Amherst, MA 01003, USA
bsilva@cs.umass.edu

Abstract

Machine learning models are often designed to maximize a primary goal, such as accuracy. However, as these models are increasingly used to inform decisions that affect people's lives or well-being, it is often unclear what the real-world repercussions of their deployment might be—making it crucial to understand and manage such repercussions effectively. Models maximizing user engagement on social media platforms, e.g., may inadvertently contribute to the spread of misinformation and content that deepens political polarization. This issue is not limited to social media—it extends to other applications where machine learning-informed decisions can have real-world repercussions, such as education, employment, and lending. Existing methods addressing this issue require prior knowledge or estimates of analytical models describing the relationship between a classifier's predictions and their corresponding repercussions. We introduce THEIA, a novel classification algorithm capable of optimizing a primary objective, such as accuracy, while providing high-confidence guarantees about its potential repercussions. Importantly, THEIA solves the open problem of providing such guarantees based solely on existing data with observations of previous repercussions. We prove that it satisfies constraints on a model's repercussions with high confidence and that it is guaranteed to identify a solution, if one exists, given sufficient data. We empirically demonstrate, using real-life data, that THEIA can identify models that achieve high accuracy while ensuring, with high confidence, that constraints on their repercussions are satisfied.

1 Introduction

Machine learning (ML) models are widely applied to real-life tasks, ranging from high-stakes applications such as lending, hiring, and criminal sentencing, to everyday applications, such as product recommendations and personalized advertisements. These models are often designed with the primary goal of maximizing the accuracy of their predictions. However, in applications where such models inform decisions that have an impact on people's lives or well-being, it is crucial to effectively manage the potential *real-life repercussions* of their deployment.

Models designed to maximize user engagement on social media, for example, have been shown to contribute to political polarization and the spread of misinformation [32, 15]. This issue extends beyond

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

social media, affecting various applications where ML-informed decisions have real-world repercussions, such as in education, employment, and lending. Previous work addressing this problem requires access to analytical models describing the relationship between a classifier's predictions and its repercussions [48, 27]. Zhu et al. [48], for example, assumes access to a model of how posts influence user opinions in order to optimize post-recommendation strategies that minimize polarization. However, this relationship is often difficult to characterize analytically. Designing classification algorithms that can mitigate repercussions without knowledge of such models remains an open problem.

We introduce THEIA, a novel classification algorithm that addresses this open problem. THEIA does not require analytical models of the repercussions of a model's predictions. Instead, it operates under the weaker assumption of having access *only* to data containing observations of the repercussions of a previously deployed classifier.

Social media example. As a running example, consider a social media platform using a classification algorithm to aid in predicting which posts to present to a user to maximize engagement. These predictions may have unintended repercussions, such as increasing political polarization. Taking repercussions into account is important—the platform may need (or be required) to maximize user engagement while managing the possible repercussions of the posts it presents. This could mean imposing constraints on metrics related to political polarization (e.g., changes in engagement or sentiment shift) after posts are presented to users.[1] Unfortunately, existing methods for this problem require analytical models of how social media posts affect political polarization. Constructing such models—e.g., through causal modeling techniques—is challenging: many complex factors (e.g., social, economic) influence the impact of the presented posts on political polarization. THEIA, by contrast, works without a model describing the relationship between predictions and their repercussions: it can ensure that constraints on the possible repercussions are satisfied with high confidence, provided the social media platform can collect data regarding the political polarization resulting from the posts presented (or recommended) by a previously deployed classifier.

Contributions. We present THEIA, a novel method capable of managing the repercussions caused by a model when the analytical relationship between model predictions and their repercussions is not known. We prove that **1)** the probability that THEIA returns a solution that satisfies constraints on possible repercussions is at least $(1-\delta)$, where δ is a user-specified confidence level; and **2)** THEIA is consistent; intuitively, it identifies and returns a model that satisfies the constraints (if one exists), with high confidence as more data is observed. We empirically analyze THEIA's performance in two real-life settings, while varying both the amount of training data and the amount of repercussions that a classifier's predictions have. We show that THEIA can identify accurate solutions while ensuring, with high confidence, that constraints on its potential repercussions are satisfied.

2 Problem formulation

THEIA's goal is to identify a high-accuracy predictive model while ensuring, with high confidence, that the *repercussions* of its deployment satisfy user-specified constraints. THEIA is designed to achieve this using *only* existing data collected during the deployment of a previous model, including empirically observed repercussions of its decisions. Concretely, we assume access to a dataset where each i^{th} data point contains a feature vector X_i, a label Y_i, and a prediction $\widehat{Y}_i^{\beta}$ made by a previously deployed stochastic classifier β. We call β the *behavior model*, defined as $\beta(x,\hat{y}) := \Pr(\widehat{Y}_i^{\beta}=\hat{y} | X_i=x)$. We assume β was trained by the same user who now seeks to improve it, so β is known, by construction, in our setting.

Consider a model trained to predict the type of content that will maximize a user's time spent on a platform. These predictions can be used to decide which posts to show to a particular user. One possible *repercussion* of deploying this model could be, e.g., how many times that user later interacted with extremist content online—a downstream effect of the content they were shown based on the model's predictions. More generally, we define a *repercussion* as a real-valued, instance-specific quantity that can be empirically observed after deploying a model and using its predictions to make decisions. Let R_i^{β} be the real-valued random variable representing the repercussion associated with the i^{th} instance, observed after the model β makes a prediction $\widehat{Y}_i^{\beta}$. In the example above, R_i^{β} is the empirical observation of how many times the i^{th} user interacted with extremist content after being

[1] For more examples of important real-life applications where THEIA can be applied, see Appendix A.

shown posts selected based on the model's prediction $\widehat{Y}_i^{\beta}$. We adopt the convention (w.l.o.g.) that smaller values of R_i^{β} correspond to more favorable repercussions. We append R_i^{β} to each data point and define the dataset as a sequence of n independent and identically distributed (i.i.d.) data points: $D{:=}(X_i, Y_i, \widehat{Y}_i^{\beta}, R_i^{\beta})_{i=1}^{n}$. We denote an arbitrary data point in D by suppressing the subscripts.

Our goal is to construct a classification algorithm that takes as input D and outputs a new model π_θ that is as accurate as possible while enforcing high-confidence constraints on repercussions.[2] If deploying a model results in repercussions that satisfy all user-specified constraints, such repercussions are referred to as ***acceptable***, and the model is deemed ***repercussion-aware***. The form of the new model is $\pi_\theta(x, \hat{y}) \coloneqq \Pr(\widehat{Y}^{\pi_\theta}{=}\hat{y}|X{=}x)$, where π_θ is parameterized by a vector $\theta \in \Theta$ (e.g., the weights of a neural network), for some feasible set Θ, and where $\widehat{Y}^{\pi_\theta}$ is the prediction made by π_θ given X. Like R_i^{β}, $R_i^{\pi_\theta}$ is the empirically observed repercussion if the model outputs the prediction $\widehat{Y}_i^{\pi_\theta}$.

Quantifying repercussions. As stated previously, we assume the repercussions of interest are measurable. Specifically, we assume there are k *repercussion objectives*, $g_j : \Theta \to \mathbb{R}, j \in 1, \ldots, k$, that take as input the parameters θ of a classifier and return a real-valued measurement of its repercussions. We adopt the convention (w.l.o.g.) that a classifier is repercussion-aware iff $g_j(\theta) \leq 0$ for all j. To simplify notation, we first investigate the setting with a single repercussion objective (i.e., $k = 1$), and later show how to enforce multiple repercussion objectives (Algorithm 3).

In this work, we investigate repercussion objectives aimed at ensuring that the repercussions of a new model π_θ will not exceed a threshold τ. Specifically, we consider cases where each repercussion objective is based on a conditional expected value of the form

$$g(\theta) \coloneqq \mathbf{E}[R^{\pi_\theta}|c(X, Y)] - \tau, \tag{1}$$

where $\tau \in \mathbb{R}$ is a tolerance and $c(X, Y)$ is a Boolean conditional; see below for more details. Appendix D shows how THEIA can enforce repercussion objectives beyond this form. These include, e.g., high-confidence guarantees that repercussions are approximately equal across different demographic groups, as well as objectives that, instead of controlling the conditional expected value, enforce high-confidence constraints on the repercussion variance, median, or conditional value at risk.

To help interpret (1), consider a platform aiming to maximize user engagement while mitigating polarization. Assume that sentiment scores are used as a proxy for polarization. Users are often segmented into groups based on factors such as age or political affiliation; $c(X, Y)$, in this case, could be a group membership variable. The platform wishes to ensure that the sentiment resulting from predictions made by a new model π_θ will be lower than τ. R^{π_θ} is the sentiment score after showing a post to a user[3] and τ is the repercussion threshold the platform does not want to exceed. This could be application-specific (e.g, $\tau{=}0.3$ implies moderate sentiment/polarization); τ could also be the current average sentiment of a particular group: $\tau{=}\frac{1}{n_1}\sum_{d=1}^{n} R_d^{\beta}[\![c(X, Y){=}1]\!]$, where $[\![\cdot]\!]$ is the Iverson bracket and n_1 is the number of people in the group. Requiring that the sentiment induced by π_θ is less than τ can be modeled by a constraint $\mathbf{E}[R^{\pi_\theta}|c(X, Y){=}1] \leq \tau$. This can be expressed in the form (1) by rewriting it as $g(\theta){=}\mathbf{E}[R^{\pi_\theta}|c(X, Y){=}1] - \tau$. Then, $g(\theta) \leq 0$ iff the future sentiment (impacted by π_θ's predictions) is at most τ. Additional constraints can be added to include other groups.

Algorithmic properties of interest. We wish to ensure that $g(\theta) \leq 0$ since this implies that θ (the model returned by a classification algorithm) has acceptable repercussions. However, this is often not possible, as it requires extensive prior knowledge of how predictions influence repercussions. Instead, we aim to design an algorithm that is capable of reasoning about its confidence that $g(\theta) \leq 0$ using only available data. That is, we wish to construct a classification algorithm, a, where $a(D) \in \Theta$ is the model returned by a when given dataset D as input, that satisfies constraints of the form

$$\Pr(g(a(D)) \leq 0) \geq 1 - \delta, \tag{2}$$

where $\delta \in (0, 1)$ is the admissible probability of returning a model with unacceptable repercussions. Algorithms satisfying (2) are called Seldonian [42]. Notice it may be impossible to enforce all constraints simultaneously or there may be insufficient data to satisfy the constraints with the required confidence. THEIA reasons about its own uncertainty and determines when this is the case.

[2] Our algorithm works with other performance objectives, not just accuracy. See Appendix D for a detailed discussion of how THEIA can address other settings and objectives.

[3] Many indicators of polarization have been proposed [7, 45, 31]. We use sentiment scores as one example.

In such situations, it proactively notifies the user that no repercussion-aware solution can be provided, returning "No Solution Found" (NSF), rather than producing a model it does not trust.[4] This can be achieved by letting $\texttt{NSF} \in \Theta$ and $g(\texttt{NSF}) = 0$.

Our goal, then, is to design a classification algorithm with two properties: **1)** the algorithm satisfies (2) and **2)** the algorithm is *consistent*, i.e., if a nontrivial repercussion-aware model exists, the probability that the algorithm returns a solution other than NSF converges to one as the amount of training data increases. In Section 4, we prove that our algorithm, THEIA, satisfies both properties.

3 Enforcing repercussion constraints

Recall that a repercussion-aware algorithm must ensure with high confidence that the repercussion objective satisfies $g(\theta) \leq 0$, where θ is the returned model and $g(\theta) = \mathbf{E}[R^{\pi_\theta}|c(X,Y)] - \tau$. Since THEIA only has access to data collected from a previously deployed model, β, the only available samples are of R^β—the repercussions of β's predictions. Below, we show *(1)* how one can construct i.i.d. estimates of R^{π_θ} using samples collected using β; *(2)* how confidence intervals can be used to derive high-confidence upper bounds on $g(\theta)$; and *(3)* the pseudocode for our algorithm.

Deriving estimates of repercussions. Recall that the distribution of empirically observed repercussions in D results from β's predictions. However, our goal is to evaluate the repercussions of a different model, π_θ. This is challenging: given empirically observed repercussions from predictions of a previously deployed model β, how to estimate the repercussions if π_θ were used instead? One solution is to run π_θ on held-out data. However, this would only produce predictions $\widehat{Y}^{\pi_\theta}$, not their corresponding repercussions; i.e., each sample's repercussion would still be in terms of β, not π_θ.

We solve this problem using off-policy evaluation methods from the RL literature, which use data collected under one decision-making model to estimate what would have happened under a different one. Specifically, we use importance sampling [34] to obtain a new random variable, $\hat{R}^{\pi_\theta}$, constructed using data from β, such that $\hat{R}^{\pi_\theta}$ is an unbiased estimator of R^{π_θ} under the standard assumptions of importance sampling (see Appendix B). For each instance in D, the estimator $\hat{R}^{\pi_\theta}$ weights the observed repercussions R^β based on how likely the prediction $\widehat{Y}^\beta$ is under π_θ. If π_θ would make the label $\widehat{Y}^\beta$ more likely, R^β is given a larger weight; otherwise, R^{π_θ} is given a smaller weight. Formally, the estimator is $\hat{R}^{\pi_\theta}{=}\pi_\theta(X,\widehat{Y}^\beta)/\beta(X,\widehat{Y}^\beta)R^\beta$.[5] Theorem 1 establishes that this estimator is unbiased.

Theorem 1. *$\hat{R}^{\pi_\theta}$ is an unbiased estimator of R^{π_θ}:* $\mathbf{E}[\hat{R}^{\pi_\theta}|c(X,Y)]{=}\mathbf{E}[R^{\pi_\theta}|c(X,Y)]$. **Proof.** *See Appendix B.* □

Bounds on repercussions. We now discuss how to derive high-confidence upper bounds on $g(\theta)$ using unbiased estimates of $g(\theta)$ and confidence intervals. While different confidence intervals for the mean can be used, we consider Student's t-test [40] and Hoeffding's inequality [20]. Consider a vector of m i.i.d. samples $(z_i)_{i=1}^m$ of a random variable Z; let the sample mean be $\bar{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$, the sample standard deviation be $\sigma(Z_1,...,Z_m){=}\sqrt{\frac{1}{m-1}\sum_{i=1}^m (Z_i - \bar{Z})^2}$, and $\delta \in (0,1)$ be a confidence level.

Property 1. *If $\sum_{i=1}^m Z_i$ is normally distributed, then* $\Pr\left(\mathbf{E}[Z_i] \geq \bar{Z} - \frac{\sigma(Z_1,...,Z_m)}{\sqrt{m}} t_{1-\delta,m-1}\right) \geq 1{-}\delta$, *where $t_{1-\delta,m-1}$ is the $1{-}\delta$ quantile of the Student's t distribution with $m{-}1$ degrees of freedom.* **Proof.** *See the work of Student [40].* □

[4] This is particularly important if one of the groups lacks sufficient data to satisfy its corresponding constraint with high confidence. In such cases, THEIA abstains from returning a model and instead outputs NSF, rather than producing a model it cannot trust. This built-in safeguard prevents the algorithm from favoring well-represented groups at the expense of those with less data, helping to avoid unintended disparities in performance or treatment.

[5] Importance sampling estimators, in general, may suffer from high variance. In our setting, two properties prevent this in practice. First, although importance sampling can require data exponential in the horizon, our classification setting involves a single decision per instance (predicting a label), making variance independent of horizon length. Second, THEIA actively rejects candidates that diverge too far from the current model: such candidates produce large importance ratios, which yield wide confidence intervals for the estimated repercussions and cause them to fail THEIA's repercussion-awareness test (Alg. 1, lines 4–9). Thus, models that could lead to high variance are naturally filtered out during the search since they induce unreliable repercussion estimates.

Property 1 can be used to obtain a high-confidence upper bound for the mean of Z: $U_{\texttt{ttest}}(Z_1, ..., Z_m) \coloneqq \bar{Z} + \frac{\sigma(Z_1,...,Z_m)}{\sqrt{m}} t_{1-\delta,m-1}$. Let $\hat{g}$ be a vector of i.i.d. and unbiased estimates of $g(\theta)$. Once computed (using importance sampling), these are provided to $U_{\texttt{ttest}}$ to derive a high-confidence upper bound on $g(\theta)$: $\Pr(\mathbf{E}[\hat{R}^{\pi_\theta}|c(X,Y)] - \tau \leq U_{\texttt{ttest}}(\hat{g})) \geq 1-\delta$. Our strategy for deriving high-confidence upper bounds for repercussion objectives is general and other confidence intervals can be used. Student's t-test may be used and holds exactly if the distribution of $\sum Z_i$ is normal. In Appendix C we describe a bound based on Hoeffding's inequality [20], which replaces the normality assumption with the weaker assumption that $\hat{g}$ is bounded, resulting in a different upper bound, $U_{\texttt{Hoeff}}$.

Complete algorithm. Algorithm 1 provides pseudocode for THEIA. It has three main steps. First, the dataset D is split between D_c and D_f (line 1). In the second step, *candidate selection* (line 3), D_c is used to find and train a model, called the *candidate solution*, θ_c. The `cost` of a candidate solution (line 3) is computed by Algorithm 2 (detailed in Appendix G), which leverages the high-confidence upper bounds introduced above.[6] In the *repercussion-awareness test* (lines 4–9), D_f is used to compute unbiased estimates of $g(\theta_c)$ using the importance sampling method described above. These estimates are used to calculate a $(1-\delta)$-confidence upper bound, U, on $g(\theta_c)$, using Hoeffding's inequality or Student's t-test (line 8). Then, U is used to decide whether θ_c or `NSF` is returned (line 9).

Algorithm 1 THEIA

Input: **1)** $D = \{(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)\}_{i=1}^n$; **2)** confidence level δ; **3)** tolerance value τ; **4)** behavior model β; and **5)** $\texttt{Bound} \in \{\texttt{Hoeff}, \texttt{ttest}\}$.
Output: Model θ_c or `NSF`.

1: $D_c, D_f \leftarrow \texttt{partition}(D)$;
2: $n_{D_f} = \texttt{length}(D_f)$; $\quad \hat{g} \leftarrow \langle\,\rangle$
3: $\theta_c \leftarrow \arg\min_{\theta\in\Theta}\ \texttt{cost}(\theta, D_c, \delta, \tau, \beta, \texttt{Bound}, n_{D_f})$
4: **for** $j \in \{1, ..., n_{D_f}\}$ **do**
5: Let $(X_j, Y_j, \widehat{Y}_j^\beta, R_j^\beta)$ be the j^{th} data point in D_f
6: **if** $c(X_j, Y_j)$ is `True` **then** $\hat{g}$.append$\left(\frac{\pi_{\theta_c}(X_j, \widehat{Y}_j^\beta)}{\beta(X_j, \widehat{Y}_j^\beta)} R_j^\beta - \tau\right)$ **end if**
7: **end for**
8: **if** Bound is `Hoeff` **then** $U = U_{\texttt{Hoeff}}(\hat{g})$ **else if** Bound is `ttest` **then** $U = U_{\texttt{ttest}}(\hat{g})$ **end if**
9: **if** $U \geq 0$ **then return** `NSF` **else return** θ_c

4 Theoretical results

This section shows that **1)** THEIA is guaranteed to satisfy the probabilistic constraints defined in (2); and **2)** given reasonable assumptions about the repercussion objectives, THEIA is consistent. Recall we wish to compute confidence intervals to bound $g(\theta_c)$,[7] where θ_c is the model returned by candidate selection. We assume that the requirements related to Student's t-test (Property 1) or Hoeffding's inequality (Appendix C) are satisfied. Let $\text{Avg}(Z) = \frac{1}{n_Z}\sum_{i=1}^{n_Z} Z_i$ be the average of a size n_Z vector Z.

Assumption 1. *If* `Bound` *is* `Hoeff`*, then for all $j \in \{1, ..., k\}$, each estimate in $\hat{g}_j$ (in Algorithm 3) is bounded in some interval $[a_j, b_j]$. If* `Bound` *is* `ttest`*, then each Avg($\hat{g}_j$) is normally distributed.*

Theorem 2. *Let $(g_j)_{j=1}^k$ be a sequence of repercussion objectives, where $g_j : \Theta \to \mathbb{R}$, and let $(\delta_j)_{j=1}^k$ be a corresponding sequence of confidence levels, where each $\delta_j \in (0, 1)$. If Assumption 1 holds, then for all $j \in \{1, ..., k\}$, $\Pr(g_j(a(D)) \leq 0) \geq 1 - \delta_j$.* **Proof.** *See Appendix E.* □

THEIA satisfies Theorem 2 if the solutions it produces satisfy (2), i.e., if $\forall j \in \{1, ..., k\}$, $\Pr(g_j(a(D)) \leq 0) \geq 1 - \delta_j$, where a is Algorithm 3. Because Algorithm 3 is an extension of Algorithm 1 to multiple constraints, it suffices to show that Theorem 2 holds for Algorithm 3. Next, we show that THEIA is consistent: when a repercussion-aware model exists, the probability that THEIA returns a solution other than `NSF` converges to 1 as the amount of training data goes to infinity.

[6] THEIA mitigates potential high-variance issues by penalizing candidate models that diverge substantially from β, as these yield wide confidence intervals unlikely to pass the repercussion-awareness test (App. G). Our experimental results are consistent with this — we did not observe high-variance issues, even in low-data regimes.

[7] THEIA works with any confidence intervals for the mean.

Theorem 3 (Consistency guarantee)**.** *If Assumptions 1–4 hold (see Appendix F for 2–4), then* $\lim_{n\to\infty} \Pr(a(D) \neq \texttt{NSF}, g(a(D)) \leq 0) = 1$. **Proof**. *We extend the proof strategy of Metevier et al. [29] from the contextual bandit setting to the supervised learning setting with high-confidence constraints. The detailed proof for this theorem is provided in Appendix F.* □

Theorem 3 relies on mild assumptions (Assumptions 2–4): **1)** the function used to evaluate models is smooth; **2)** at least one repercussion-aware model exists that is not on the acceptable-repercussions boundary; and **3)** a model's sample performance converges to it expected value given enough data.

5 Empirical evaluation

We empirically investigate three research questions: **RQ1:** Does THEIA enforce repercussion constraints with high probability, while existing algorithms do not? **RQ2:** What is the cost (e.g., in terms of accuracy) of enforcing such constraints? **RQ3:** How does THEIA perform when predictions have little influence on the repercussions of interest relative to factors outside of the model's control?

In our experiments, we examine two real-life use cases of THEIA focused on ensuring *fair repercussions*, i.e., that deploying the model will benefit different demographic groups approximately equally.[8] Recent research has shown that ML models can impact different demographic groups differently [5]. In our first experiment (EXP-1), a classifier makes predictions about whether youth in the U.S. foster care system are likely to get a job. These predictions may affect a person's life if, e.g., they influence financial aid decisions. EXP-1 uses two data sources from the National Data Archive on Child Abuse and Neglect [1], which include financial, educational, and well-being data on youth over time and during their transition from foster care to adulthood. The feature vector X contains five attributes related to the job and educational status of a person in foster care, including their race. The goal is to predict a binary label, Y, denoting whether a person has a full-time job after leaving the program. The behavior model, β, is a logistic regression classifier. We explore our research questions in a setting where classifier predictions may have different levels of repercussions. To do so, we consider a parameterized definition of a measurable repercussion. Let R_i^ψ be the observed repercussion on person i's life if a classifier ψ outputs the prediction $\widehat{Y}_i^\psi$ given X_i. Here, $\psi(x, \hat{y}) \coloneqq \Pr(\widehat{Y}_i^\psi{=}\hat{y} | X_i{=}x)$. We define R_i^ψ as

$$R_i^\psi = -\begin{cases} \alpha\widehat{Y}_i^\psi + (1-\alpha)\ \mathcal{N}(2, 0.5) & \text{if } X_i^r = 0 \\ \alpha\widehat{Y}_i^\psi + (1-\alpha)\ \mathcal{N}(1, 1) & \text{if } X_i^r = 1, \end{cases} \tag{3}$$

where X_i^r is the race of person i, and α regulates whether a model's repercussions are strongly affected by its predictions (*high prediction-repercussion dependency*) or if they have little to no repercussions (*low prediction-repercussion dependency*). As α goes to zero, predictions have no repercussions. We vary α from 0 to 1 in increments of 0.1, and ψ is defined as the behavior model, β. Intuitively, Equation 3 captures the idea that if a youth is predicted to get a job, they are more likely to receive financial aid—leading to less severe observed repercussions. That is, R_i^ψ decreases when $\widehat{Y}_i^\psi = 1$. Notice that both the mean and variance of the repercussion resulting from a given prediction vary by race. This reflects the fact that youth from different racial backgrounds may experience systematically different outcomes, e.g., due to structural or social biases. This form aligns with the empirical findings of Chetty et al. [10], who show, using real-world data, that the distribution of financial outcomes (i.e., repercussions) differs across racial groups even when starting from similar conditions (e.g., having the same prediction $\widehat{Y}_i^\psi$). Notice that Equation 3 characterizes the setting we investigate and is used solely to generate challenging, parameterized datasets with varying levels of prediction-repercussion dependency. THEIA does *not* have access to it, nor does it depend on or assume its particular form.

In our second experiment (EXP-2), a bank's lending decisions are informed by a classifier predicting repayment success. Such decisions can have repercussions on clients' lives, affecting their financial well-being, savings rate, or debt-to-income ratio after a lending decision. EXP-2 uses real-life financial information for 250,000 clients who requested loans [44]. The feature vector X contains attributes such as a client's age, monthly income, debt-to-income ratio, number of open loans, number of dependents, and various delinquency measures. The goal is to predict a binary label Y, denoting

[8] Our notion of *repercussion fairness* differs qualitatively from standard *static fairness* definitions. Static fairness metrics rely on performance metrics like false positive rate and can be determined by evaluating a model on a validation set. *Repercussion fairness*, by contrast, is more general and requires deploying the model and empirically observing/measuring its repercussions.

whether a client will miss any payments by more than 90 days. THEIA's primary objective is to approve loan requests for clients unlikely to miss payments while accounting for the possible repercussions of lending decisions across age groups (attribute X^r): younger individuals who are often not yet financially stable, for example, may experience a greater impact on their financial well-being from receiving a new loan than more senior individuals. In this experiment, R_i^ψ is the observed repercussion on client i's financial well-being if a classifier ψ outputs the prediction $\widehat{Y}_i^\psi$ given X_i. For a thorough discussion, see Appendix H. Importantly, in EXP-2 there exists a non-linear relationship between predictions and their repercussions due to the Law of Diminishing Marginal Utility of Income: as income increases (e.g., through an approved loan), the benefits of further increasing it decrease non-linearly [28]. Furthermore, this experiment models a challenging scenario where distribution shift causes the repercussions of ML models to change between the deployment of the baseline classifier (behavior model) THEIA aims to improve upon and the time when THEIA searches for a repercussion-aware model. Specifically, it models a scenario where significant societal changes have altered the relationship between predictions and their repercussions, reflecting, for example, how loan decisions may impact clients differently over time. This allows us to empirically evaluate THEIA's robustness in cases where statistical properties of the data change after the training data was collected.

In both EXP-1 and EXP-2, we wish to guarantee that the repercussions of the new model, π_θ, are smaller (i.e., less severe) than those of the current model, β. In particular, THEIA's goal is to identify an accurate predictive model that is repercussion-aware, ensuring its repercussions are not detrimental to a particular race (EXP-1) or age group (EXP-2). To do so, we define two repercussion objectives, g_0 and g_1. Let $t \in \{0, 1\}$ and $g_t(\theta) := \mathbf{E}[R^{\pi_\theta} | X^r = t] - \tau_t$, where $\tau_t = \frac{1}{n_t} \sum_{d=1}^{n} R_d^\beta [\![X_d^r = t]\!]$ is the average observed repercussion caused by β on people of race (or age group) $X^r = t$ and where $n_t = \sum_{d=1}^{n} [\![X_d^r = t]\!]$. The confidence level δ_t for all objectives is 0.1.

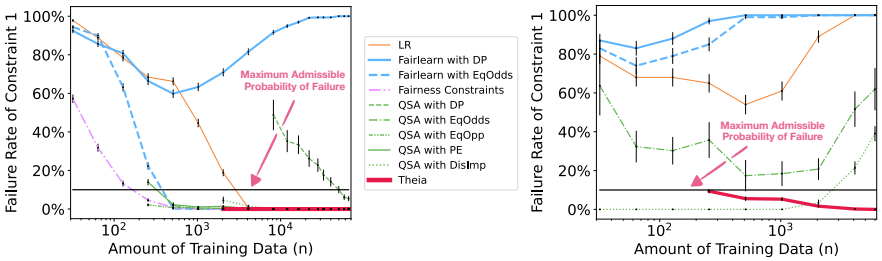


Figure 1: Failure rates w.r.t. a repercussion constraint [**Left**: EXP-1; **Right**: EXP-2]. Black lines show the maximum admissible probability of failing to enforce the constraints ($\delta_0 = \delta_1 = 10\%$).

RQ1: Managing repercussions. We first investigate if THEIA can effectively satisfy repercussion constraints while existing algorithms fail to do so. A key novelty of THEIA is its ability to ensure repercussion fairness using only data from a previously deployed classifier, without relying on analytical models describing how predictions affect repercussions. While model-based approaches exist, ours is, to the best of our knowledge, the first capable of satisfying repercussion constraints in the fully model-free setting. As such, there are no direct baselines that operate solely on data collected from a previously deployed classifier. We therefore compare THEIA to well-established state-of-the-art model-free fairness methods that make assumptions aligned with ours and that also rely only on observed data. In particular, we compare with **1)** Fairlearn [2], **2)** Fairness Constraints [46], and **3)** quasi-Seldonian algorithms (QSA) [42]. We consider five classic fairness constraints: demographic parity (DP), equalized odds (EqOdds), disparate impact (DisImp), equal opportunity (EqOpp), and predictive equality (PE) [11, 14, 18].[9] We also compare THEIA to a simple baseline: logistic regression (LR).

We first examine how often each algorithm fails to satisfy repercussion constraints. Let the *failure rate* be the probability an algorithm returns a solution violating a constraint. We measure this by evaluating the solutions returned by each algorithm on a larger dataset not used during training. In EXP-1, we also investigate how varying the prediction-repercussion dependency, α, influences failure rates. We focus on one representative experiment with $\alpha = 0.9$ due to space constraints. Similar behavior is observed for other α values. See Appendix H for full results and implementation details.

Figure 1 shows each algorithm's failure rate, as a function of training data size, in each experiment (EXP-1 and EXP-2). We computed failure rates and standard errors over 500 trials. Figure 1 depicts

[9] We task each competing method with enforcing the constraints analyzed in those methods' original papers.

failure rates with respect to a representative repercussion constraint; complete results are in Appendix H. Notice that the solutions returned by THEIA *always* satisfy both repercussion constraints.[10] This is consistent with THEIA's theoretical guarantees, which ensure with high probability that solutions it returns satisfy all constraints. Other methods, by contrast, either **1)** *always fail* to satisfy both constraints; or **2)** fail to satisfy one constraint while only occasionally satisfying the other.

RQ1: *Our experiments show that, with high probability,* THEIA *identifies repercussion-aware models satisfying all user-specified constraints, while alternative methods fail to do so.*

RQ2: The cost of managing repercussions. Depending on the problem, there may be a trade-off between satisfying repercussion constraints and optimizing accuracy. In Appendix D, we show how THEIA can satisfy repercussion constraints while bounding accuracy loss. We now investigate the impact that enforcing repercussion constraints has on accuracy. Figure 2 shows the accuracy of classifiers returned by different algorithms as a function of n in EXP-1; the results of EXP-2 (Appendix H) follow the same pattern. We bounded accuracy loss via an additional constraint requiring that returned models have an accuracy of at least 75%. Under low-data regimes, THEIA's accuracy is approximately 90%, whereas competing methods (with no repercussions guarantees) exhibit a slightly better performance above 90%. As n increases, THEIA's accuracy reaches that of the other techniques. Importantly, although competing methods may occasionally achieve slightly higher accuracy than ours, they *consistently return models that violate the constraints on the repercussions*. THEIA, by contrast, ensures that *all constraints are satisfied* with high probability and *always* returns models with accuracy above the specified threshold (see Figure 1).

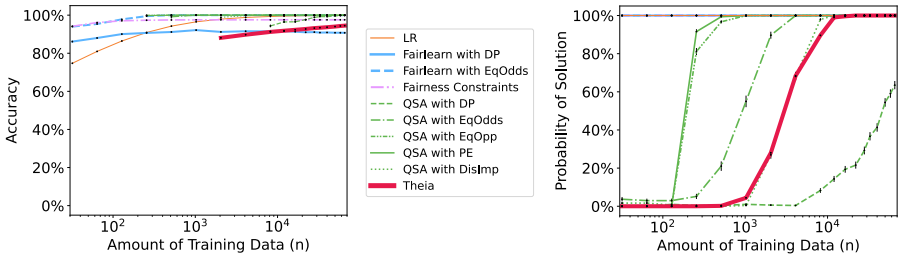


Figure 2: [**Left:** EXP-1] Accuracy of the models returned by algorithms (subject to different constraints) as a function of n. [**Right:** EXP-1] Probability that these algorithms return a solution.

A trade-off might also exist between the amount of available training data and the confidence that models satisfying all constraints can be identified. Recall that some methods (including THEIA) may choose not to return a solution if they cannot confidently ensure that all constraints are satisfied. We now study how often each method identifies a candidate solution as a function of n. Figure 2 shows that THEIA starts returning models that satisfy all constraints, with high confidence, when given a number of samples corresponding to just 3.1% of the available data, and it is capable of *consistently* returning models (with 90% probability) when given a number of samples corresponding to just 12.5% of the available data. As n increases, the probability of THEIA returning models increases rapidly. Although Fairlearn, Fairness Constraints, and LR return a model regardless of the amount of training data, these models *never satisfy both constraints* (see Figure 1). QSA often returns candidate models with less training data than THEIA; these models, however, also *fail to satisfy both constraints* simultaneously.

RQ2: *While there is a cost to managing a model's repercussions depending on the setting,* THEIA *succeeds in its primary objective: ensuring, with high confidence, that repercussion constraints are satisfied without requiring unreasonable amounts of data, and while limiting accuracy loss.*

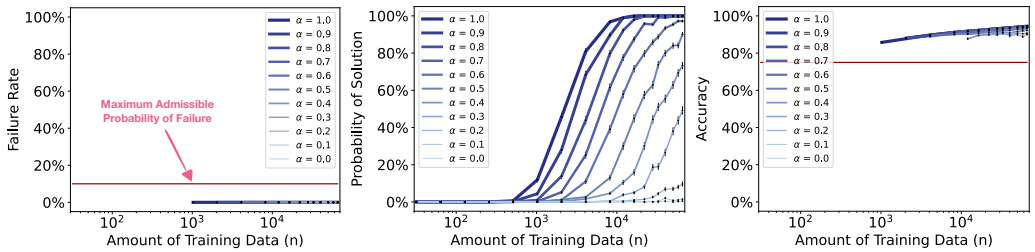
RQ3: Varying prediction-repercussion dependency. Finally, we investigate THEIA's performance (in terms of failure rate, probability of returning solutions, and accuracy) in settings with varied levels of prediction-repercussion dependency. These include challenging cases where predictions have little influence on the observed repercussions, relative to other factors outside of the model's control.

We first study THEIA's failure rate for different values of α. Figure 3a shows that THEIA *never* returns solutions that fail to satisfy the repercussion constraints, independent of α, confirming empirically that

[10] THEIA returns a model only when it can ensure, with high confidence, that all constraints are satisfied. In EXP-1, this requires just $\approx 1.5\%$ of the available data; in EXP-2, it requires only $\approx 0.1\%$.

its high-probability guarantees hold in settings with a wide range of qualitatively different observed repercussion characteristics. Next, we investigate how often THEIA identifies and returns a model for various values of α. If predictions have little to no influence on observed repercussions (i.e., low α), it becomes difficult to distinguish minor repercussions from noise—making it harder for THEIA to confidently identify models that satisfy all repercussion constraints. As expected, this reduces the likelihood of returning a solution (see Figure 3b). THEIA returns models for all $\alpha > 0$ given sufficient data, but this probability decreases as α approaches zero. This is by design: THEIA naturally handles noisy repercussions by reasoning about its uncertainty—noise inflates confidence intervals, making it less likely that it will have sufficient confidence in its predictions. It returns a model only if enough data is available to offset this effect; otherwise, it proactively warns the user that no repercussion-aware model can be identified. Lastly, we investigate how the amount of prediction-repercussion dependency affects the accuracy of THEIA's models. Figure 3c shows model accuracy for various values of α, as a function of n. The accuracy trade-off is more evident when THEIA must satisfy challenging repercussion objectives, as α approaches zero. In these cases, accuracy decreases from 95% to 90%. Importantly, however, even though a trade-off exists, our method remains successful at bounding accuracy while ensuring, with high confidence, that *all* constraints on the model's repercussions are satisfied.

RQ3: *Our experiments confirm that* THEIA *performs well in a wide range of settings, with various levels of prediction-repercussion dependency. Although managing repercussions may impact accuracy and the probability of finding solutions, these unavoidable trade-offs do* ***not*** *affect* THEIA*'s high confidence guarantees. In our experiments, all returned solutions satisfy both constraints.*



(a) Failure rate. (b) Prob. of returning a solution. (c) Accuracy.

Figure 3: [EXP-1] THEIA's performance under various levels of prediction-repercussion dependency.

6 Related work

Our paper addresses the *side effects problem*, i.e., how to ensure that predictive models are effective without causing unintended consequences [4, 26, 37]. Below, we discuss relevant work related to ours.

Counterfactual prediction. Counterfactual prediction has been studied in the off-policy policy evaluation (OPE) [34, 41] and causal inference literature [33]. THEIA uses importance sampling to construct unbiased estimators of a model's repercussions without having to deploy it. This type of counterfactual reasoning aligns with how importance sampling is used in OPE. Recent work has shown that counterfactual modeling is required if ML-informed decisions act as risk-mitigating interventions [27, 12]. Methods addressing this setting include, e.g., techniques to manage long-term fairness; importantly, they assume a known analytical relationship between predictions and long-term impact. Constructing these models is challenging: complex factors (social, economic, etc.) may influence, e.g., how financial decisions affect various demographic groups. We, by contrast, investigate a novel *model-free* method that reasons about the repercussions of predictive models based *only* on existing data.

Dynamics modeling in content delivery settings. A substantial body of literature focuses on modeling the dynamics of systems that optimize personalized content delivery. Social media platforms, for instance, use algorithms to determine which posts to present to users to maximize engagement [30, 47, 35, 16, 6]. These algorithms typically assume access to a model of how a recommendation system (RS) may affect user preferences and content exposure [24, 23, 8]. Tommasel et al. [43] model the dynamics of echo chambers to construct a friend recommendation system that mitigates echo chamber effects by enhancing recommendation diversity. We extend the state of the art by introducing a model-free technique that, with high confidence, mitigates repercussions such as the exacerbation of existing echo chambers. Carroll et al. [8] investigate a setting similar to ours. They optimize an RS so it does not result in a manipulative or undesirable influence on user preferences. This is achieved by training

a dynamics model of how user preferences evolve under new RS policies, and using it to identify "safe" RS policies. THEIA, by contrast, can identify repercussion-aware models using a model-free OPE approach and is supported by strong formal guarantees on user-defined notions of repercussions.

Fairness over time. Liu et al. [27] showed that classifiers' predictions that appear fair with respect to static fairness criteria can negatively impact the long-term wellness of the community it aims to protect. Many works addressed the challenge of identifying classifiers that optimize measures of fairness over time [21, 22, 27, 13, 19]. Existing methods assume access to a model relating decisions and their repercussions on different populations [27, 47, 17]. We showed that it is possible to achieve this same goal using only existing data, without having to learn models—which is often challenging or infeasible. Appendix D discusses how this setting may be mapped to our mathematical framework.

Seldonian algorithms. THEIA extends the existing body of work on Seldonian algorithms [42], which provide high-confidence guarantees on user-defined metrics of interest, such as fairness and safety, and have shown strong performance in real-world applications [42, 29]. They also provide a straightforward way for users to define multiple constraints on metrics of interest [42]. THEIA is the first classification Seldonian algorithm capable of providing high-confidence guarantees that constraints on the repercussions of deploying a predictive model will be satisfied.

7 Conclusions

We introduced THEIA, a novel approach to mitigating the real-life repercussions of deploying ML models. Unlike existing methods, it does not rely on complex analytical models describing the relationship between a classifier's predictions and their repercussions. THEIA, by contrast, is the first method that can provably provide high-confidence guarantees on a model's repercussions using only available data. Importantly, it can reason about its own uncertainty: if there is insufficient data to satisfy all constraints with the required confidence, it proactively notifies the user that no safe solution can be provided, rather than returning a model it does not trust. A promising direction for future research is extending THEIA to settings with complementary assumptions about the statistical properties of training data, such as those where the relationship between predictions and repercussions evolves over time.

8 Acknowledgments

This work is supported by the National Science Foundation under grant no. CCF-2210243.

References

[1] Administration on Children, Youth & Families. National Data Archive on Child Abuse and Neglect (NDACAN). 2021. URL `https://www.ndacan.acf.hhs.gov/`.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *35th International Conference on Machine Learning*, pages 60–69, 2018.

[3] Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[5] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 145(1):30–56, 2021.

[6] Elisabetta Biondi, Chiara Boldrini, Andrea Passarella, and Marco Conti. Dynamics of opinion polarization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[7] Dennies Bor, Benjamin Seiyon Lee, and Edward J Oughton. Quantifying polarization across political groups on key policy issues using sentiment analysis. *arXiv preprint arXiv:2302.07775*, 2023.

[8] Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pages 2686–2708. PMLR, 2022.

[9] Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

[10] Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783, 2020.

[11] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[12] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

[13] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[15] Miriam Fernandez, Alejandro Bellogín, and Iván Cantador. Analysing the effect of recommendation algorithms on the spread of misinformation. In *Proceedings of the 16th ACM Web Science Conference*, 2024.

[16] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990.

[17] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 445–453, 2021.

[18] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 3315–3323, 2016.

[19] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *36th International Conference on Machine Learning*, pages 2692–2701, 2019.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[21] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[22] Lily Hu and Yiling Chen. Welfare and distributional impacts of fair classification. *arXiv preprint arXiv:1807.01134*, 2018.

[23] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.

[24] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. Preference amplification in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 805–815, 2021.

[25] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[26] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

[27] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *35th International Conference on Machine Learning*, pages 3150–3158, 2018.

[28] A. Marshall. *Principles of Economics: An Introductory Volume*. Macmillan, 1920.

[29] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip Thomas. Offline contextual bandits with high probability fairness guarantees. In *Advances in Neural Information Processing Systems 32*, 2019.

[30] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*, pages 369–378, 2018.

[31] Christopher Musco, Indu Ramesh, Johan Ugander, and R Teal Witter. How to quantify polarization in models of opinion dynamics. *arXiv preprint arXiv:2110.11981*, 2021.

[32] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web*, 17(4):1–26, 2023.

[33] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[34] Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *18th International Conference on Machine Learning*, pages 417–424, 2001.

[35] Anton V Proskurnikov and Roberto Tempo. A tutorial on modeling and analysis of dynamic social networks. part i. *Annual Reviews in Control*, 43:65–79, 2017.

[36] Ingo Rechenberg and Manfred Eigen. Evolutionsstrategie: Optimierung technischer systeme nach prinzipien der biologischen evolution. *Frommann-Holzboog Stuttgart*, 1973.

[37] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. A multi-objective approach to mitigate negative side effects. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 354–361, 2021.

[38] Hans-Paul Schwefel. Numerische optimierung von computer-modellen mittels der evolutionsstrategie. 1977.

[39] Pranab K. Sen and Julio M. Singer. *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, Boca Raton, FL, 1993.

[40] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[41] Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.

[42] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019.

[43] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. I want to break free! recommending friends from outside the echo chamber. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 23–33, 2021.

[44] Will Cukierski Credit Fusion. Give Me Some Credit Dataset, 2011. URL `https://www.kaggle.com/competitions/GiveMeSomeCredit`.

[45] Muheng Yang, Xidao Wen, Yu-Ru Lin, and Lingjia Deng. Quantifying content polarization on twitter. In *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*, pages 299–308. IEEE, 2017.

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *20th Artificial Intelligence and Statistics*, pages 962–970, 2017.

[47] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3):7–11, 2020.

[48] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. *Advances in Neural Information Processing Systems*, 34: 2072–2084, 2021.

Appendix

A Other motivating examples

In our paper we discussed two important real-life examples where our method could be applied. First, we introduced a motivating problem where a social media platform wished to take into account the repercussions of presenting users with different posts. Secondly, in our empirical evaluation, we considered the repercussions of providing financial aid to youth in foster care. Here, we discuss three additional examples of possible applications of THEIA in real-life settings:

- Consider a university that has a 1-on-1 tutoring program. However, the university has limited resources and cannot offer tutoring for all students. To select which students should participate in the program, the university's decision is based on GPA predictions. This can have repercussions in students' lives: receiving tutoring (or not) can influence the chances of a student graduating from college.
- Consider a police department deciding which crime prevention strategy to use in each district of a city, based on predictions about crime recidivism. This could have repercussions in the form of, e.g., changes in the average incarceration rate after this decision.
- Assume that medical decisions are influenced by predictions of whether a person qualifies for high-risk care management. These predictions may have a repercussion on a person's health; e.g., the severity of chronic illnesses after receiving the treatment.

B Proof of Theorem 1

Proof. At a high level, we start with $\mathbf{E}[\hat{R}^{\pi_\theta} \mid c(X,Y)]$ and apply a series of transformations, standard probabilistic identities, and general properties of the setting (see below) to arrive at $\mathbf{E}[R^{\pi_\theta} \mid c(X,Y)]$. To simplify notation, throughout this proof we let $C = c(X,Y)$. Also, for any random variable Z, let $\mathrm{supp}(Z)$ denote the support of Z (e.g., if Z is discrete, then $\mathrm{supp}(Z) = \{z : \Pr(Z=z) > 0\}$). To begin, we substitute the definition of $\hat{R}^{\pi_\theta}$ and expand this expression using the definition of expected value:

$$\mathbf{E}[\hat{R}^{\pi_\theta}|C] = \mathbf{E}\left[\frac{\pi_\theta(X,\widehat{Y}^\beta)}{\beta(X,\widehat{Y}^\beta)}R^\beta \middle| C\right] \tag{4}$$

$$= \sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^\beta,R^\beta)} \Pr(X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, R^\beta{=}r|C)\frac{\pi_\theta(x,\hat{y})}{\beta(x,\hat{y})}r. \tag{5}$$

Using the chain rule repeatedly, we can rewrite the joint probability in (5) as follows:

$$\Pr(X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, R^\beta{=}r|C) \tag{6}$$

$$= \Pr(R^\beta{=}r|X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, C)\Pr(X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}|C) \tag{7}$$

$$= \Pr(R^\beta{=}r|X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, C)\Pr(\widehat{Y}^\beta{=}\hat{y}|X{=}x, Y{=}y, C)\Pr(X{=}x, Y{=}y|C). \tag{8}$$

Furthermore, recall that our algorithm addresses a classification problem. In standard machine learning settings, the predicted label depends only on the input features provided to the model. As a result, conditioning the label's probability on any additional variables not seen by the model does not change that probability. That is, $\Pr(\widehat{Y}^\beta{=}\hat{y}|X{=}x, Y{=}y, C) = \Pr(\widehat{Y}^\beta{=}\hat{y}|X{=}x)$, which is the definition of $\beta(x,\hat{y})$.

We perform this substitution and simplify by canceling out the β terms:

$$\mathbf{E}[\hat{R}^{\pi_\theta}|C] = \sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^\beta,R^\beta)} \Pr\left(R^\beta{=}r|X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, C\right)\beta(x,\hat{y})\Pr\left(X{=}x, Y{=}y|C\right)\frac{\pi_\theta(x,\hat{y})}{\beta(x,\hat{y})}r \tag{9}$$

$$= \sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^\beta,R^\beta)} \Pr\left(R^\beta{=}r|X{=}x, Y{=}y, \widehat{Y}^\beta{=}\hat{y}, C\right)\Pr\left(X{=}x, Y{=}y|C\right)\pi_\theta(x,\hat{y})r. \tag{10}$$

Note that $\pi_\theta(x,\hat{y})$ can be rewritten as $\Pr(\widehat{Y}^{\pi_\theta}{=}\hat{y}|X{=}x,Y{=}y,C)$. Using the multiplication rule of probability, we can combine this term with the $\Pr(X{=}x,Y{=}y|C)$ term in (10) to obtain the joint probability $\Pr(X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}{=}\hat{y}|C)$. Furthermore, notice that in the setting we investigate, predictions' repercussions do not depend on *how* a prediction was made. That is, it does not matter, e.g., if a neural network or a random forest determined which posts to show to a user. Formally, this means that $\forall x,y,\hat{y},r,\ \Pr(R^{\beta}{=}r|X{=}x,Y{=}y,\widehat{Y}^{\beta}{=}\hat{y}){=}\Pr(R^{\pi_\theta}{=}r|X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}{=}\hat{y})$. For this reason, we can substitute $\Pr(R^{\beta}{=}r|X{=}x,Y{=}y,\widehat{Y}^{\beta}{=}\hat{y},C)$ for $\Pr(R^{\pi_\theta}{=}r|X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}{=}\hat{y},C)$. We substitute these terms into (10) and apply the multiplication rule of probability once more:

$$\mathbf{E}[\hat{R}^{\pi_\theta}|C] = \sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^{\beta},R^{\beta})} \Pr(R^{\pi_\theta}{=}r|X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}{=}\hat{y},C)\Pr(X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}=\hat{y}|C)r \tag{11}$$

$$= \sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^{\beta},R^{\beta})} \Pr(X{=}x,Y{=}y,\widehat{Y}^{\beta}{=}\hat{y},\hat{R}^{\pi_\theta}{=}r|C)r. \tag{12}$$

Finally, we make the standard assumption in the importance sampling literature that predictions made by π_θ can, in principle, be evaluated using the available data. This means that predictions that are possible under π_θ (the model being evaluated) have a non-zero probability of occurring under β (the classifier used to generate training data). In other words, for all x and y, $\pi_\theta(x,y)>0$ implies that $\beta(x,y)>0$. This assumption is trivially satisfied in our setting: we consider predictive models $\theta\in\Theta$ that naturally satisfy this condition, such as standard stochastic classifiers that assign non-zero probability to all outputs (e.g., Softmax layers in neural networks). Under this assumption, $\mathrm{supp}(\widehat{Y}^{\pi_\theta})\subseteq\mathrm{supp}(\widehat{Y}^{\beta})$, and so $\mathrm{supp}(R^{\pi_\theta})\subseteq\mathrm{supp}(R^{\beta})$. So, we can rewrite (12) as

$$\sum_{(x,y,\hat{y},r)\in\mathrm{supp}(X,Y,\widehat{Y}^{\pi_\theta},R^{\pi_\theta})} \Pr(X{=}x,Y{=}y,\widehat{Y}^{\pi_\theta}{=}\hat{y},\hat{R}^{\pi_\theta}{=}r|C)r. \tag{13}$$

By the definition of expectation, this is equivalent to $\mathbf{E}\left[R^{\pi_\theta}|C\right]$. Therefore, we have shown that $\mathbf{E}[\hat{R}^{\pi_\theta}|C]{=}\mathbf{E}[R^{\pi_\theta}|C]$. □

C Bounds on repercussions using Hoeffding's inequality

This section focuses on how one can use unbiased estimates of $g(\theta)$ together with Hoeffding's inequality [20] to derive high-confidence upper bounds on $g(\theta)$. Given a vector of m i.i.d. samples $(Z_i)_{i=1}^m$ of a random variable Z, let $\bar{Z}=\frac{1}{m}\sum_{i=1}^m Z_i$ be the sample mean, and let $\delta\in(0,1)$ be a confidence level.

Property 2 (Hoeffding's Inequality)**.** *If* $\Pr(Z\in[a,b])=1$*, then*

$$\Pr\left(\mathbf{E}[Z_i]\geq\bar{Z}-(b-a)\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\geq 1-\delta. \tag{14}$$

Proof. *See the work of Hoeffding [20].* □

Property 2 can be used to obtain a high-confidence upper bound for the mean of Z:

$$U_{\texttt{Hoeff}}(Z_1,Z_2,...,Z_m)\coloneqq\bar{Z}+(b{-}a)\sqrt{\frac{\log(1/\delta)}{(2m)}}. \tag{15}$$

Let $\hat{g}$ be a vector of i.i.d. and unbiased estimates of $g(\theta)$. Once these are procured (using importance sampling as described in Section 3), they can be provided to $U_{\texttt{Hoeff}}$ to derive a high-confidence upper bound on $g(\theta)$:

$$\Pr\left(\mathbf{E}\left[\hat{R}^{\pi_\theta}\middle|c(X,Y)\right]-\tau\leq U_{\texttt{Hoeff}}(\hat{g})\right)\geq 1-\delta. \tag{16}$$

Notice that using Hoeffding's inequality to obtain the upper bound requires the assumption that $\hat{g}$ is bounded.

D Extensions of THEIA

In this section we discuss how THEIA can be extended to provide similar high-confidence guarantees for the regression setting and for repercussion objectives beyond the form assumed in (1).

D.1 Repercussion-awareness guarantees in the regression setting

In our problem setting, we study repercussion-awareness in the classification setting, in which the labels $\widehat{Y}$ produced by a model are discrete. However, our method can also be applied in the regression setting, where a (stochastic) regression model produces *continuous* predictions $\widehat{Y}$, instead of discrete labels. To use THEIA in this setting, one may adapt Algorithm 2 so that it uses a loss function suitable for regression; e.g., sample mean squared error. Furthermore, notice that the importance sampling technique described in Section 3 is still applicable in the regression setting, requiring only minor changes so that it can be used in such a continuous setting. In particular, the importance sampling technique we described can be adapted by replacing summations with integrals, probability mass functions with probability density functions, and probabilities with probability densities. By doing so, all results presented in our work (e.g., regarding the unbiasedness of the importance sampling estimator) carry to the continuous case. Notice, finally, that in order to apply THEIA in the regression setting, the behavior model, β, and the new candidate model, π_θ, must be *stochastic* regression models—this is similar to the assumption we made when addressing the classification setting (see the discussion in Section 2).

D.2 Alternative definitions of repercussion objectives

Until now, we have assumed that the repercussion objectives take the form of (1). However, this can be restrictive when the user's notion of repercussion-awareness requires a different definition of repercussions. Below, we discuss how users of THEIA may construct other repercussion objectives, and how our formulation (shown in (1)) is related to the definitions introduced in the work of Liu et al. [27].

Connections to delayed-impact fairness and the work of Liu et al. [27]. Consider the lending scenario described in Liu et al. [27], in which a bank's objective is to maximize loan repayments, and its lending decisions are informed by a classifier that predicts repayment success. For simplicity, the population of loan applicants consists of two mutually exclusive groups, A and B (e.g., based on race or gender). The repercussions of lending decisions are multi-fold—payment defaults not only reduce the bank's profit, but worsen the financial situation of the borrowers, whereas successful repayments lead to profit for the bank and an increase in the borrowers' financial welfare.

In their work, Liu et al. [27] define long-term improvement and long-term stagnation as repercussion objectives that do not cause long-term harm. Specifically, they use $\Delta\mu_j$ to represent the difference in repercussions (such as changes in financial welfare) for group $j \in \{A, B\}$ between a previously deployed model and a new model. If $\Delta\mu_j < 0$, then the new model has resulted in long-term harm.

To enforce long-term fairness in our framework, we assume that the training dataset provided has data instances of the form $(X, Y, T, \hat{Y}^\beta, R^\beta)$, where X is a real-valued vector describing information about a loan applicant, Y indicates whether a loan should have been approved, $T \in \{A, B\}$ is the group indicator, $\hat{Y}^\beta$ is the prediction a previous classifier β made given X, and R^β is the resulting change in credit score of the individual, which Liu et al. [27] use as a proxy for financial well-being in their work. To map this example to our framework, we can set τ to be group T's average credit score under the current model (i.e., under the behavior model, β) and $\mathbf{E}\left[R^{\pi_\theta} | T = j\right]$ to be the expected credit score of group j under the new model, π_θ. Then, $\Delta\mu_j = \mathbf{E}\left[R^{\pi_\theta} | T = j\right] - \tau$.

Enforcing general definitions of repercussion objectives To enforce repercussion objectives beyond (1), the importance sampling technique introduced in Section 3 can be combined with techniques presented in the work of Metevier et al. [29]. As a concrete example, consider the lending scenario described in the previous section. Assume that instead of long-term improvement or stagnation, the bank is interested in ensuring (with high probability) that the repercussions of a classifier's predictions are approximately equal for loan applicants in groups A and B. This can be represented by the repercussion objective $g_{\text{fair}}(\theta) = \left|\mathbf{E}\left[R^{\pi_\theta} | T = A\right] - \mathbf{E}\left[R^{\pi_\theta} | T = B\right]\right| - \epsilon$.

To satisfy (2) with the repercussion objective g_{fair} in place of g, the techniques in this paper and the bound-propagation methods introduced by Metevier et al. [29] can be combined. At a high-level, THEIA would first compute (as before) unbiased estimates of $\mathbf{E}\left[R^{\pi_\theta}|T=A\right]$ and $\mathbf{E}\left[R^{\pi_\theta}|T=B\right]$ using the importance sampling technique described in Section 3. Then, it would use the bound-propagation methods introduced by Metevier et al. [29] to obtain high-confidence upper bounds on $g_{\text{fair}}(\theta)$.

The discussion above corresponds to just one example of how to deal with alternative repercussion objectives; in this particular example, g_{fair}. The same general idea and techniques can also be used to tackle alternative repercussion objectives that users of THEIA may be interested in.[11] For example, in settings where one wishes to minimize repercussions (rather than merely constrain them), a repercussion threshold τ can be chosen based on a target value deemed sufficiently low—thus ensuring that any returned model also achieves near-optimal repercussion values. All other parts of the algorithm would remain the same—e.g., the algorithm would still split the dataset into two, identify a candidate solution, and check whether it passes the repercussion-awareness test.

Beyond conditional expectation. In Section 2, we assume that g is defined in terms of the conditional expected value of the repercussion objectives. However, other forms of repercussion awareness might be more appropriate for different applications. For example, conditional value at risk [25] might be appropriate for risk-sensitive applications, and the median might be relevant for applications with noisy data [3]. Chandak et al. [9] introduce off-policy evaluation methods that produce estimates and high-confidence bounds for different distributional parameters of interest, including value at risk, conditional value at risk, variance, median, and interquantile range. These techniques can be combined with ours to obtain high-confidence upper bounds for metrics other than the conditional expected value of R^{π_θ}.

E Proof of Theorem 2

This section proves Theorem 2, which is restated below.

Theorem 2: Let $(g_j)_{j=1}^k$ be a sequence of repercussion objectives, where $g_j : \Theta \to \mathbb{R}$, and let $(\delta_j)_{j=1}^k$ be a corresponding sequence of confidence levels, where each $\delta_j \in (0,1)$. If Assumption 1 holds, then for all $j \in \{1,...,k\}$,

$$\Pr(g_j(a(D)) \leq 0) \geq 1 - \delta_j. \tag{17}$$

In what follows, we consider the same conditions as in Theorem 1, which we restate below for completeness:

- Conditioning the probability of a label on variables not observed by the classification model does not affect that probability.
- The repercussions of a prediction do not depend on *how* a prediction was made—it does not matter, e.g., if a neural network or a random forest determined which posts to show to a user.
- Predictions made by π_θ can, in principle, be evaluated using the available data. Specifically, predictions that are possible under π_θ (the model being evaluated) have a non-zero probability of occurring under β (the classifier used to generate training data).

We first provide three lemmas that will be used when proving Theorem 2.

Lemma 1. *Let $\hat{g}_j$ be the estimates of g constructed in Algorithm 3, and let D_{f_c} be a subdataset of D_f such that a data point $(X, Y, \widehat{Y}^\beta, R^\beta)$ is only in D_{f_c} if $c(X,Y)$ is true. Then, for all $\theta \in \Theta$, the elements in $\hat{g}_j$ are $i.i.d.$ samples from the conditional distribution of $\hat{g}_j$ given $c(X,Y)$.*

[11] This statement holds assuming that the repercussion objective of interest satisfies the requirements for the bound-propagation technique to be applicable; for example, that the repercussion objective can be expressed using elementary arithmetic operations (e.g., addition and subtraction) over *base variables* for which we know unbiased estimators [29]. In the case of the repercussion objectives discussed in this paper, we can obtain unbiased estimates of the relevant quantities using importance sampling, as discussed in Section 3.

Proof. To obtain $\hat{g}_j$, each data point in D_{f_c} is transformed into an estimate of $g(\theta)$ using the importance sampling estimate $\frac{\pi_\theta(X,\widehat{Y}^\beta)}{\beta(X,\widehat{Y}^\beta)}R^\beta - \tau$ (Algorithm 3, lines 5–8). Since each element of $\hat{g}_j$ is computed from a single data point in D_{f_c}, and the points in D_{f_c} are conditionally independent given $c(X,Y)$, it follows that each element of $\hat{g}_j$ is conditionally independent given $c(X,Y)$. So, each element of $\hat{g}_j$ can be viewed as an i.i.d. sample from the conditional distribution of $\hat{g}_j$ given $c(X,Y)$. □

Lemma 2. *Let $\hat{g}_j$ be the estimates of g constructed in Algorithm 3. It follows from Theorem 1 that for all $\theta \in \Theta$, each element in $\hat{g}_j$ is an unbiased estimate of $g_j(\theta)$.*

Proof. We begin by considering the expected value of any element in $\hat{g}_j$:

$$\begin{aligned}
\mathbf{E}\left[\frac{\pi_\theta(X,\widehat{Y}^\beta)}{\beta(X,\widehat{Y}^\beta)}R^\beta\middle| c(X,Y)-\tau\right] =& \mathbf{E}\left[\frac{\pi_\theta(X,\widehat{Y}^\beta)}{\beta(X,\widehat{Y}^\beta)}R^\beta\middle| c(X,Y)\right]-\tau && (18)\\
=& \mathbf{E}\left[\hat{R}^{\pi_\theta}\middle| c(X,Y)\right]-\tau && (19)\\
=& \mathbf{E}\left[R^{\pi_\theta}|c(X,Y)\right]-\tau && (20)\\
=& g_j(\theta). && (21)
\end{aligned}$$

Expression (20) follows from Theorem 1. Therefore, for all $\theta \in \Theta$, the elements of $\hat{g}_j$ are unbiased estimates of $g_j(\theta)$. □

Let θ_c be the model returned by candidate selection in Algorithm 3 (line 2), and let U_j be the value of U at iteration j of the for loop (lines 4–10).

Lemma 3. *If Lemmas 1 and 2 hold, then the upper bounds U_j calculated in Algorithm 3 satisfy $\forall j \in \{1,...,k\}$, $\Pr(g_j(\theta_c) > U_j) \leq \delta_j$.*

Proof. We begin by noting that by Lemma 1, the data points used to construct each $(1-\delta_j)$-probability bound, i.e., the data points in each $\hat{g}_j$, are (conditionally) i.i.d. Because $\theta_c \in \Theta$, by Lemma 2 we know that each element in $\hat{g}_j$ is an unbiased estimate of $g_j(\theta_c)$. Therefore, Hoeffding's inequality or Student's t-test can be applied to random variables that are (conditionally) i.i.d.[12] and unbiased estimates of $g_j(\theta_c)$. Moreover, under Assumption 1, when `Bound` is `Hoeff`, the requirements of Hoeffding's inequality are satisfied (Property 2), and when `Bound` is `ttest`, the requirements of Student's t-test are satisfied (Property 1). Therefore, the upper bounds calculated in Algorithm 3 satisfy $\Pr(g_j(\theta_c) > U_j) \leq \delta_j$. □

Proof of Theorem 2

Proof. To show Theorem 2, we prove the contrapositive, i.e., $\forall j \in \{1,...,k\}, \Pr(g_j(a(D)) > 0) \leq \delta_j$.

Consider the event $\forall j \in \{1,...,k\}, g_j(a(D)) > 0$. When this event occurs, it is always the case that $a(D) \neq \texttt{NSF}$ (by definition, $g(\texttt{NSF}) = 0$). That is, a nontrivial solution was returned by the algorithm, and for all j, $U_j \leq 0$ (line 11 of Algorithm 3). Therefore, (22) (shown below) holds.

$$\begin{aligned}
\Pr(g_j(a(D)>0) =& \Pr(g_j(a(D)) > 0, U_j \leq 0) && (22)\\
\leq& \Pr(g_j(a(D)) > U_j) && (23)\\
=& \Pr(g_j(\theta_c) > U_j) && (24)\\
\leq& \delta_j. && (25)
\end{aligned}$$

Expression (23) is a result of the fact that the joint event in (22) implies the event $(g_j(a(D)) > U_j)$. We substitute θ_c for $a(D)$ in (24) because the event $\forall j \in \{1,...,k\}, g_j(a(D)) > 0$ implies that a nontrivial solution, or a solution that is not `NSF`, was returned: $a(D) = \theta_c$. Lastly, (25) follows from Lemma 3. This implies that $\Pr(g_j(a(D) > 0) \leq \delta_j \; \forall j \in \{1,...,k\}$, completing the proof.

□

[12] Samples that are conditionally i.i.d. given some event E can be viewed as i.i.d. samples from the conditional distribution. Applying the confidence intervals to these samples provides high-confidence bounds on the *conditional* expected value given the event E, which is precisely what we aim to bound.

F Proof of Theorem 3

This section proves Theorem 3, restated below. We build upon the proof strategy introduced by Metevier et al. [29], extending it from the contextual bandit setting to the supervised learning setting with high-confidence constraints. Extending their proof to our setting involves the following changes:

1. Changes related to the output of the function used to calculate the utility of a model: Metevier et al. [29] consider a utility function that returns the sample reward of a policy. Instead, our utility function (Algorithm 2) outputs the sample loss of a model.
2. Changes due to the form of the constraints: The form of our repercussion constraint differs from the more general form of the constraints considered by Metevier et al. [29]. This results in a simplified argument that our algorithm is consistent.

We present the complete proof, with these changes incorporated, below.

Theorem 3: If Assumptions 1–4 hold, then $\lim_{n\to\infty} \Pr(a(D) \neq \texttt{NSF},\ g(a(D))\leq 0\big)=1$.

We begin by providing definitions and assumptions necessary for presenting our main result. To simplify notation, we assume that there exists only a single constraint and note that the extension of this proof to multiple constraints is straightforward. As before, we consider the same conditions as in Theorem 1, which we restate below for completeness:

- Conditioning the probability of a label on variables not observed by the classification model does not affect that probability.
- The repercussions of a prediction do not depend on *how* a prediction was made—it does not matter, e.g., if a neural network or a random forest determined which posts to show to a user.
- Predictions made by π_θ can, in principle, be evaluated using the available data. Specifically, predictions that are possible under π_θ (the model being evaluated) have a non-zero probability of occurring under β (the classifier used to generate training data).

Recall that the logged data, D, is a random variable. To further formalize this notion, let (Ω, Σ, p) be a probability space on which all relevant random variables are defined, and let $D_n : \Omega \to \mathcal{D}$ be a random variable, where $\mathcal{D}$ is the set of all possible datasets and $D_n = D_c \cup D_f$. We will discuss convergence as $n \to \infty$. $D_n(\omega)$ is a particular sample of the entire set of logged data with n data points, where $\omega \in \Omega$.

Definition 1 (Piecewise Lipschitz continuous)**.** *We say that a function $f : M \to \mathbb{R}$ on a metric space (M, d) is piecewise Lipschitz continuous with Lipschitz constant K and with respect to a countable partition, $\{M_1, M_2, ...\}$, of M if f is Lipschitz continuous with Lipschitz constant K on all metric spaces in $\{(M_i, d)\}_{i=1}^{\infty}$.*

Definition 2 (δ-covering)**.** *If (M, d) is a metric space, a set $X \subseteq M$ is a δ-covering of (M, d) if and only if $\max_{y\in M} \min_{x\in X} d(x, y) \leq \delta$.*

Let $\hat{c}(\theta, D_c)$ denote the output of a call to Algorithm 2, and let $c(\theta) \coloneqq \ell_{\max} + g(\theta)$. The next assumption ensures that c and $\hat{c}$ are piecewise Lipschitz continuous. Notice that the δ-covering requirement is straightforwardly satisfied if Θ is countable or $\Theta \subseteq \mathbb{R}^m$ for any positive natural number m.

Assumption 2. *The feasible set of policies, Θ, is equipped with a metric, d_Θ, such that for all $D_c(\omega)$ there exist countable partitions of Θ, $\Theta^c = \{\Theta_1^c, \Theta_2^c, ...\}$, and $\Theta^{\hat{c}} = \{\Theta_1^{\hat{c}}, \Theta_2^{\hat{c}}, ...\}$, where $c(\cdot)$ and $\hat{c}(\cdot, D_c(\omega))$ are piecewise Lipschitz continuous with respect to Θ^c and $\Theta^{\hat{c}}$ respectively with Lipschitz constants K and $\hat{K}$. Furthermore, for all $i \in \mathbb{N}_{>0}$ and all $\delta > 0$ there exist countable δ-covers of Θ_i^c and $\Theta_i^{\hat{c}}$.*

Intuitively, Assumption 2 states that *(1)* the cost function used to evaluate classifiers is smooth: similar classifiers have similar costs/performances; and *(2)* each classifier can be described by a set of real-valued parameters, as is the case with all parametric supervised learning algorithms.

Next, we assume that a repercussion-aware model, $\theta^\star$, exists such that $g(\theta^\star)$ is not precisely on the boundary of acceptable repercussions. This can be satisfied by models that are arbitrarily close to such boundary.

Assumption 3. *There exists an $\epsilon > \xi$ and a $\theta^\star \in \Theta$ such that $g(\theta^\star) \leq -\epsilon$.*

Intuitively, Assumption 3 states that the space of classifiers is not degenerate: at least one repercussion-aware model exists such that if we perturb its parameters infinitesimally, it would not arbitrarily no longer satisfy the repercussion objective. Next, we assume that the sample loss, $\hat{\ell}$, converges almost surely to ℓ, the actual expected loss.

Assumption 4. $\forall \theta \in \Theta$, $\hat{\ell}(\theta, D_c) \xrightarrow{a.s.} \ell(\theta)$.

Intuitively, Assumption 4 states that the sample performance of a classifier converges to its true expected performance given enough data. This is similar to the usual assumption, e.g., in the regression setting, that a model's sample Mean Squared Error (MSE) converges to its true MSE given sufficient examples.

We prove Theorem 3 by building up properties that culminate with the desired result, starting with a variant of the strong law of large numbers:

Property 3 (Khintchine Strong Law of Large Numbers)**.** *Let $\{X_\iota\}_{i=1}^{\infty}$ be independent and identically distributed random variables. Then $(\frac{1}{n}\sum_{i=1}^{n} X_\iota)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbf{E}[X_1]$, if $\mathbf{E}[X_1]$ exists, i.e., $\frac{1}{n}\sum_{i=1}^{n} X_\iota \xrightarrow{a.s.} \mathbf{E}[X_1]$.*

Proof. See Theorem 2.3.13 of Sen and Singer [39]. □

Next, we show that the average of the estimates of $g(\theta)$ converges almost surely to $g(\theta)$:

Property 4. *If Lemmas 1 and 2 hold, then $\forall \theta \in \Theta, \mathrm{Avg}(\hat{g}) \xrightarrow{a.s.} g(\theta)$.*

Proof. Recall that if Lemmas 1 and 2 hold, estimates in $\hat{g}$ are i.i.d. and each estimate in $\hat{g}$ is an unbiased estimate of $g(\theta)$. Also, recall that if $n_{\hat{g}}$ is the number of elements in $\hat{g}$, $\mathrm{Avg}(\hat{g}) \coloneqq \frac{1}{n_{\hat{g}}}\sum_{i=1}^{n_{\hat{g}}} \hat{g}_i$. Then, by Property 3 we have that $\mathrm{Avg}(\hat{g}) \xrightarrow{\text{a.s.}} g(\theta)$. □

In this proof, we consider the set $\bar{\Theta} \subseteq \Theta$, which contains all models that are not repercussion-aware, and some that are repercussion-aware but fall beneath a certain threshold: $\bar{\Theta} \coloneqq \{\theta \in \Theta : g(\theta) > -\xi/2\}$. At a high level, we will show that the probability that the candidate model, θ_c, viewed as a random variable that depends on the candidate dataset D_c, satisfies $\theta_c \notin \bar{\Theta}$ converges to one as $n \to \infty$, and then that the probability that θ_c is returned also converges to one as $n \to \infty$.

First, we will show that the upper bounds U^+ (constructed in candidate selection, i.e., Algorithm 2) and U (constructed in the repercussion-awareness test, i.e., Algorithm 1) converge to $g(\theta)$ for all $\theta \in \Theta$. To clarify notation, we write $U^+(\theta, D_c)$ and $U(\theta, D_f)$ to emphasize that each depends on θ and the datasets D_c and D_f, respectively.

Property 5. *If Assumption 1 holds, then it follows from Property 4 that for all $\theta \in \Theta$, $U^+(\theta, D_c) \xrightarrow{a.s.} g(\theta)$ and $U(\theta, D_f) \xrightarrow{a.s.} g(\theta)$.*

Proof. Given Assumption 1, Hoeffding's inequality and Student's t-test construct high-confidence upper bounds on the mean by starting with the sample mean of the unbiased estimates (in our case, $\mathrm{Avg}(\hat{g})$) and then adding an additional term (a constant in the case of Hoeffding's inequality). Thus, $U(\theta, D_f)$ can be written as $\mathrm{Avg}(\hat{g}) + Z_n$, where Z_n is a sequence of random variables that converges (surely for Hoeffding's inequality, almost surely for Student's t-test) to zero. So, $Z_n \xrightarrow{\text{a.s.}} 0$, and we need only show that $\mathrm{Avg}(\hat{g}) \xrightarrow{\text{a.s.}} g(\theta)$, which follows from Property 4. We therefore have that $U \xrightarrow{\text{a.s.}} g(\theta)$.

The same argument can be used when substituting $U^+(\theta, D_c)$ for $U(\theta, D_f)$. Notice that the only difference between the method used to construct confidence intervals in the repercussion-awareness test (that is, U^+) and in Algorithm 2 (that is, U) is the multiplication of Z_n by a constant λ. This still results in a sequence of random variables that converges (almost surely for Student's t-test) to zero. □

Recall that we define $\hat{c}(\theta, D_c)$ to be the output of Algorithm 2. Below, we show that given a repercussion-aware model $\theta^\star$ and data D_c, $\hat{c}(\theta^\star, D_c)$ converges almost surely to $\ell(\theta^\star)$, the expected loss of $\theta^\star$.

Property 6. *If Assumptions 3 and 4 hold, then it follows from Property 5 that* $\hat{c}(\theta^\star, D_c) \xrightarrow{a.s.} \ell(\theta^\star)$.

Proof. By Property 5, we have that $U^+(\theta^\star) \xrightarrow{\text{a.s.}} g(\theta^\star)$. By Assumption 3, we have that $g(\theta^\star) \leq -\epsilon$. Now, let

$$A = \{\omega \in \Omega : \lim_{n\to\infty} U^+(\theta^\star, D_c(\omega)) = g(\theta^\star)\}. \tag{26}$$

Recall that $U^+(\theta^\star, D_c) \xrightarrow{\text{a.s.}} g(\theta^\star)$ means that $\Pr(\lim_{n\to\infty} U^+(\theta^\star, D_c) = g(\theta^\star)) = 1$. So, ω is in A almost surely, i.e., $\Pr(\omega \in A) = 1$. Consider any $\omega \in A$. From the definition of a limit and the previously established property that $g(\theta^\star) \leq -\epsilon$, we have that there exists an n_0 such that for all $n \geq n_0$, Algorithm 2 will return $\hat{\ell}(\theta^\star, D_c)$ (this avoids the discontinuity of the `if` statement in Algorithm 2 for values smaller than n_0).

Furthermore, we have from Assumption 4 that $\hat{\ell}(\theta^\star, D_c) \xrightarrow{\text{a.s.}} \ell(\theta^\star)$. Let

$$B = \{\omega \in \Omega : \lim_{n\to\infty} \hat{\ell}(\theta^\star, D_c(\omega)) = \ell(\theta^\star)\}. \tag{27}$$

From Assumption 4, we have that ω is in B almost surely, i.e., $\Pr(\omega \in B) = 1$, and thus by the countable additivity of probability measures, $\Pr(\omega \in (A \cap B)) = 1$.

Consider now any $\omega \in (A \cap B)$. We have that for sufficiently large n, Algorithm 2 will return $\hat{\ell}(\theta^\star, D_c)$ (since $\omega \in A$), and further that $\hat{\ell}(\theta^\star, D_c) \to \ell(\theta^\star)$ (since $\omega \in B$). Thus, for all $\omega \in (A \cap B)$, the output of Algorithm 2 converges to $\ell(\theta^\star)$, i.e., $\hat{c}(\theta^\star, D_c(\omega)) \to \ell(\theta^\star)$. Since $\Pr(\omega \in (A \cap B)) = 1$, we conclude that $\hat{c}(\theta^\star, D_c(\omega)) \xrightarrow{\text{a.s.}} \ell(\theta^\star)$. □

We have now established that the output of Algorithm 2 converges almost surely to $\ell(\theta^\star)$ for the $\theta^\star$ assumed to exist in Assumption 3. We now establish a similar result for all $\theta \in \bar{\Theta}$—that the output of Algorithm 2 converges almost surely to $c(\theta)$ (recall that $c(\theta)$ is defined as $\ell_{\max} + g(\theta)$).

Property 7. *It follows from Property 5 that for all* $\theta \in \bar{\Theta}$, $\hat{c}(\theta, D_c) \xrightarrow{a.s.} c(\theta)$.

Proof. By Property 5, we have that $U^+(\theta, D_c) \xrightarrow{\text{a.s.}} g(\theta)$. If $\theta \in \bar{\Theta}$, then we have that $g(\theta) > -\xi/2$. We now change the definition of the set A from its definition in the previous property to a similar definition suited to this property. That is, let:

$$A = \{\omega \in \Omega : \lim_{n\to\infty} U^+(\theta, D_c(\omega)) = g(\theta)\}. \tag{28}$$

Recall that $U^+(\theta, D_c) \xrightarrow{\text{a.s.}} g(\theta)$ means that $\Pr(\lim_{n\to\infty} U^+(\theta, D_c) = g(\theta)) = 1$. So, ω is in A almost surely, i.e., $\Pr(\omega \in A) = 1$. Consider any $\omega \in A$. From the definition of a limit and the previously established property that $g(\theta) > -\xi/2$, we have that there exists an n_0 such that for all $n \geq n_0$ Algorithm 2 will return $\ell_{\max} + U^+(\theta, D_c(\omega))$. By Property 5, $U^+(\theta, D_c(\omega)) \xrightarrow{\text{a.s.}} g(\theta)$. So, for all $\omega \in A$, the output of Algorithm 2 converges almost surely to $\ell_{\max} + g(\theta)$; that is, $\hat{c}(\theta, D_c(\omega)) \xrightarrow{\text{a.s.}} \ell_{\max} + g(\theta)$, and since $c(\theta) = \ell_{\max} + g(\theta)$, we therefore conclude that $\hat{c}(\theta, D_c(\omega)) \xrightarrow{\text{a.s.}} c(\theta)$. □

By Property 7 and one of the common definitions of almost sure convergence,

$$\forall \theta \in \bar{\Theta}, \forall \epsilon > 0, \Pr\Big(\lim_{n\to\infty} \inf\{\omega \in \Omega : |\hat{c}(\theta, D_n(\omega)) - c(\theta)| < \epsilon\}\Big) = 1.$$

Because Θ is not countable, it is not immediately clear that all $\theta \in \bar{\Theta}$ converge simultaneously to their respective $c(\theta)$. We show next that this is the case due to our smoothness assumptions.

Property 8. *If Assumption 2 holds, then it follows from Property 7 that* $\forall \epsilon' > 0$,

$$\Pr\Big(\lim_{n\to\infty} \inf\{\omega \in \Omega : \forall \theta \in \bar{\Theta}, |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon'\}\Big) = 1. \tag{29}$$

Proof. Let $\mathrm{C}(\delta)$ denote the union of all the points in the δ-covers of the countable partitions of Θ assumed to exist by Assumption 2. Since the partitions are countable and the δ-covers for each region are assumed to be countable, we have that $\mathrm{C}(\delta)$ is countable for all δ. Then by Property 7, for all δ, we have convergence for all $\theta \in C(\delta)$ simultaneously:

$$\forall \delta > 0, \forall \epsilon > 0, \Pr\Big(\lim_{n\to\infty} \inf\{\omega \in \Omega : \forall \theta \in C(\delta), |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon\}\Big) = 1. \tag{30}$$

Now, consider a $\theta \notin C(\delta)$. By Definition 2 and Assumption 2, $\exists\ \theta' \in \bar{\Theta}_i^c, d(\theta, \theta') \leq \delta$. Moreover, because c and $\hat{c}$ are Lipschitz continuous on $\bar{\Theta}_i^c$ and $\bar{\Theta}_i^{\hat{c}}$ (by Assumption 2) respectively, we have that $|c(\theta) - c(\theta')| \leq K\delta$ and $|\hat{c}(\theta, D_c(\omega)) - \hat{c}(\theta', D_c(\omega))| \leq \hat{K}\delta$. So, $|\hat{c}(\theta, D_c(\omega)) - c(\theta)| \leq |\hat{c}(\theta, D_c(\omega)) - c(\theta')| + K\delta \leq |\hat{c}(\theta', D_c(\omega)) - c(\theta')| + \delta(K + \hat{K})$. This means that for all $\delta > 0$:

$$\Big(\forall \theta \in \mathrm{C}(\delta), |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon\Big) \implies \Big(\forall \theta \in \bar{\Theta}, |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon + \delta(K + \hat{K})\Big).$$

Substituting this into (30), we get:

$$\forall \delta > 0, \forall \epsilon > 0, \Pr\Big(\lim_{n\to\infty} \inf\{\omega \in \Omega : \forall \theta \in \bar{\Theta},\ |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon + \delta(K + \hat{K})\}\Big) = 1.$$

Now, let $\delta \coloneqq \epsilon/(K + \hat{K})$ and $\epsilon' = 2\epsilon$. Thus, we have the following:

$$\forall \epsilon' > 0, \Pr\Big(\lim_{n\to\infty} \inf\{\omega \in \Omega : \forall \theta \in \bar{\Theta}, |\hat{c}(\theta, D_c(\omega)) - c(\theta)| < \epsilon'\}\Big) = 1.$$

□

So, given the appropriate assumptions, for all $\theta \in \bar{\Theta}$, we have that $\hat{c}(\theta, D_c(\omega)) \xrightarrow{\text{a.s.}} c(\theta)$ and that $\hat{c}(\theta^\star, D_c(\omega)) \xrightarrow{\text{a.s.}} \ell(\theta^\star)$. Due to the countable additivity property of probability measures and Property 8, we have the following:

$$\Pr\Big(\Big[\forall \theta \in \bar{\Theta}, \lim_{n\to\infty} \hat{c}(\theta, D_c(\omega)) = c(\theta)\Big],\ \Big[\lim_{n\to\infty} \hat{c}(\theta^\star, D_c(\omega)) = \ell(\theta^\star)\Big]\Big) = 1, \tag{31}$$

where $\Pr(A, B)$ denotes the joint probability of A and B.

Let H denote the set of $\omega \in \Omega$ such that (31) is satisfied. Note that $\ell_{\max}$ is defined as the value always greater than $\ell(\theta)$ for all $\theta \in \Theta$, and $g(\theta) \geq -\xi$ for all $\theta \in \bar{\Theta}$. So, for all $\omega \in H$, for sufficiently large n, candidate selection will not define θ_c to be in $\bar{\Theta}$. Since ω is in H almost surely ($\Pr(\omega \in H) = 1$), we therefore have that $\lim_{n\to\infty} \Pr(\theta_c \notin \bar{\Theta}) = 1$.

The remaining challenge is to establish that, given $\theta_c \notin \bar{\Theta}$, the probability that the repercussion-awareness test returns θ_c rather than `NSF` converges to one as $n \to \infty$. By Property 5, we have that $U(\theta_c, D_f) \xrightarrow{\text{a.s.}} g(\theta_c)$. Furthermore, by the definition of $\bar{\Theta}$, when $\theta_c \notin \bar{\Theta}$ we have that $g(\theta_c) < -\xi/2$. So, $U(\theta_c, D_f)$ converges almost surely to a value less than $-\xi/2$. Since the repercussion-awareness test returns θ_c rather than `NSF` if $U(\theta_c, D_f) \leq -\xi/4$ and $U(\theta_c, D_f)$ converges almost surely to a value less than $-\xi/2$, it follows that the probability that $U(\theta_c, D_f) \leq -\xi/4$ converges to one. Hence, given that $\theta_c \notin \bar{\Theta}$, the probability that θ_c is returned rather than `NSF` converges to one.

We therefore have that **1)** the probability that $\theta_c \notin \bar{\Theta}$ converges to one as $n \to \infty$ and **2)** given that $\theta_c \notin \bar{\Theta}$, the probability that θ_c is returned rather than `NSF` converges to one. Since $\theta_c \notin \bar{\Theta}$ implies that θ_c is repercussion-aware, these two properties imply that the probability that a repercussion-aware model is returned converges to one as $n \to \infty$.

G Full Algorithm

Algorithm 2 corresponds to the cost function used in candidate selection (line 3 of Algorithm 1). When given a candidate model, θ, to evaluate, the cost function first uses the training set, D_c, to estimate whether the candidate model is likely to pass the repercussions-awareness test. This is done by using D_c to compute an upper bound, U^+, on $g(\theta_c)$. If $U^+ \leq -\xi/4$, a small negative constant,[13]

[13]We consider $-\xi/4$, instead of 0 as the threshold in THEIA to ensure consistency. Appendix F discusses this in more detail.

Algorithm 2 `cost`

Input: **1)** the vector θ that parameterizes model π; **2)** $D_c = \{(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)\}_{i=1}^m$; **3)** confidence level δ; **4)** tolerance value τ; **5)** the behavior model β; **6)** $\texttt{Bound} \in \{\texttt{Hoeff}, \texttt{ttest}\}$; and **7)** the number of data points in D_f, denoted n_{D_f}.
Output: The cost of π.

1: $\hat{g} \leftarrow \langle\,\rangle$
2: **for** $i \in \{1, ..., m\}$ **do**
3: **if** $c(X_i, Y_i)$ is `True` **then** $\hat{g}$.append$\Big(\frac{\pi_\theta(X_i, \widehat{Y}_i^\beta)}{\beta(X_i, \widehat{Y}_i^\beta)} R_i^\beta - \tau\Big)$ **end if**
4: **end for**
5: Let $\lambda = 2$; $\quad n_{\hat{g}} = \texttt{length}(\hat{g})$
6: **if** Bound is `Hoeff` **then**
7: $a, b \leftarrow$ upper and lower bounds of g
8: $U^+ = \frac{1}{n_{\hat{g}}} \left(\sum_{\iota=1}^{n_{\hat{g}}} \hat{g}_\iota\right) + \lambda(b-a)\sqrt{\frac{\log(1/\delta)}{(2n_{D_f})}}$
9: **else if** Bound is `ttest` **then** $U^+ = \frac{1}{n_{\hat{g}}} \left(\sum_{\iota=1}^{n_{\hat{g}}} \hat{g}_\iota\right) + \lambda \frac{\sigma(\hat{g})}{\sqrt{n_{D_f}}} t_{1-\delta, n_{D_f}-1}$ **end if**
10: $\ell_{\max} = \max_{\theta' \in \Theta} \hat{\ell}(\theta', D_c)$
11: **if** $U^+ \leq -\frac{\xi}{4}$ **return** $\hat{\ell}(\theta, D_c)$ **else return** $(\ell_{\max} + U^+)$

Algorithm 3 THEIA with Multiple Constraints

Input: **1)** $D = \{(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)\}_{i=1}^n$; **2)** the number of repercussion constraints, k; **3)** a sequence of Boolean conditionals $(c_j)_{j=1}^k$ such that for $j \in \{1, ..., k\}$, $c_j(X_i, Y_i)$ indicates whether the event associated with the data point $(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)$ occurs; **4)** confidence levels $\delta = (\delta_j)_{j=1}^k$, where each $\delta_j \in (0, 1)$ corresponds to repercussion constraint g_j; **5)** tolerance values $\tau = (\tau_j)_{j=1}^k$, where each τ_j is the tolerance associated with repercussion constraint g_j; **6)** the behavior model β; and **7)** $\texttt{Bound} \in \{\texttt{Hoeff}, \texttt{ttest}\}$.
Output: Model θ_c or `NSF`.

1: $D_c, D_f \leftarrow \texttt{partition}(D)$
2: $\theta_c \leftarrow \arg\min_{\theta \in \Theta} \texttt{cost}(\theta, D_c, k, \delta, \tau, \beta, \texttt{Bound}, \texttt{length}(D_f))$
3: $U \leftarrow \langle\,\rangle$
4: **for** $j \in \{1, ..., k\}$ **do**
5: $\hat{g}_j \leftarrow \langle\,\rangle$
6: **for** $i \in \{1, ..., n\}$ **do**
7: **if** $c_j(X_i, Y_i)$ is `True` **then** $\hat{g}_j$.append$\Big(\frac{\pi_{\theta_c}(X_i, \widehat{Y}_i^\beta)}{\beta(X_i, \widehat{Y}_i^\beta)} R_i^\beta - \tau_j\Big)$ **end if**
8: **end for**
9: **if** Bound is `Hoeff` **then** U.append$(U_{\texttt{Hoeff}}(\hat{g}_j))$ **else** U.append$(U_{\texttt{ttest}}(\hat{g}_j))$ **end**
10: **end for**
11: **if** $\forall j \in \{1, ..., k\}, U_j \leq 0$ **then return** θ_c **else return** `NSF`

Algorithm 2 determines that θ is likely to pass the repercussion-awareness test, and the cost associated with the loss of θ is returned. Otherwise, the cost of θ is defined as the sum of U^+ and the maximum loss that can be obtained on dataset D_c (lines 10–11). This discourages candidate selection from returning models unlikely to pass the repercussion-awareness test.

Notice that Algorithm 2 uses importance sampling ratios between two models—the currently deployed classifier, β, and a candidate model, π_θ—to compute an unbiased estimate of the repercussion objective, g. If nothing is known in advance about the relationship between these models, the resulting ratios can be extremely small or large, leading to high-variance estimators. However, the candidate models evaluated by THEIA are *not* arbitrary. THEIA is designed to reject candidates that diverge too much from the current model: such models produce large importance ratios, which lead to wide confidence intervals and typically fail THEIA's repercussion-awareness test (lines 4–9, Algorithm 1). These candidate models, therefore, are filtered out during the search. In particular, wide confidence intervals result in large upper bounds, U^+, on the repercussion objective. These bounds are likely

Algorithm 4 `cost` with Multiple Constraints

Input: **1)** the vector θ that parameterizes the model π; **2)** $D_c = \{(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)\}_{i=1}^m$; **3)** the number of repercussion constraints, k; **4)** a sequence of Boolean conditionals $(c_j)_{j=1}^k$ such that for $j \in \{1, ..., k\}$, $c_j(X_i, Y_i)$ indicates whether the event associated with the data point $(X_i, Y_i, \widehat{Y}_i^\beta, R_i^\beta)$ occurs; **5)** confidence levels $\delta = \{\delta_j\}_{j=1}^k$, where each $\delta_j \in (0, 1)$ corresponds with constraint g_j; **6)** tolerance values $\tau = \{\tau_j\}_{j=1}^k$, where each τ_j is the tolerance associated with repercussion constraint g_j; **7)** the behavior model β; **8)** `Bound` $\in \{$`Hoeff`, `ttest`$\}$; and **9)** the number of data points in D_f, denoted n_{D_f}.

Output: The cost associated with the model π_θ.

1: **for** $j \in \{1, ..., k\}$ **do**
2: $\quad \hat{g}_j \leftarrow \langle\,\rangle$
3: $\quad$ **for** $i \in \{1, ..., m\}$ **do**
4: $\quad\quad$ **if** $c_j(X_i, Y_i)$ is `True` **then** $\hat{g}_j$.append$\left(\frac{\pi_\theta(X_i, \widehat{Y}_i^\beta)}{\beta(X_i, \widehat{Y}_i^\beta)} R_i^\beta - \tau_j\right)$ **end if**
5: $\quad$ **end for**
6: $\quad$ Let $\lambda = 2$; $\quad n_{\hat{g}_j} = $ `length`$(\hat{g}_j)$
7: $\quad$ **if** `Bound` is `Hoeff` **then**
8: $\quad\quad$ Let a, b be the lower and upper bounds of g_j
9: $\quad\quad$ $U_j^+ = \frac{1}{n_{\hat{g}_j}} \left(\sum_{\iota=1}^{n_{\hat{g}_j}} (\hat{g}_j)_\iota\right) + \lambda(b-a)\sqrt{\frac{\log(1/\delta_j)}{(2n_{D_f})}}$
10: $\quad$ **else if** `Bound` is `ttest` **then**
11: $\quad\quad$ $U_j^+ = \frac{1}{n_{\hat{g}_j}} \left(\sum_{\iota=1}^{n_{\hat{g}_j}} (\hat{g}_j)_\iota\right) + \lambda \frac{\sigma(\hat{g}_j)}{\sqrt{n_{D_f}}} t_{1-\delta_j, n_{D_f}-1}$
12: $\quad$ **end if**
13: **end for**
14: $\ell_{\max} = \max_{\theta' \in \Theta} \hat{\ell}(\theta', D_c)$
15: **if** $\forall j \in \{1, ..., k\}, U_j^+ \leq -\xi/4$ **then return** $\hat{\ell}(\theta, D_c)$
16: **else return** $\left(\ell_{\max} + \sum_{j=1}^k U_j^{\text{inflated}}\right)$

to exceed the threshold $-\frac{\xi}{4}$, causing the algorithm to assign the maximum possible cost, $\ell_{\max}$, to high-variance candidates—effectively eliminating them from consideration (line 11, Algorithm 2). In practice, thus, this mechanism serves as a way to implicitly constrain importance sampling ratios and control the variance of THEIA's estimates of the repercussion objective.

Furthermore, notice that in Algorithm 2, instead of calculating a high-confidence upper bound on $g(\theta)$ using $U_{\texttt{Hoeff}}$ or $U_{\texttt{ttest}}$, we calculate an *inflated* upper bound U^+. Specifically, we inflate the width of the confidence interval used to compute the upper bound (lines 5–9). This is to mitigate the fact that multiple comparisons are performed on the same dataset (D_c) during the search for a candidate model, which often leads candidate selection to overestimate its confidence that the model it picks will pass the repercussion-awareness test. Our choice to inflate the confidence interval in this way, i.e., considering the size of the dataset D_f used in the repercussion-awareness test and the use of scaling constant λ, is empirically driven and was first proposed for other Seldonian algorithms [42].

Algorithm 3 shows THEIA with multiple constraints. The changes relative to Algorithm 1 are relatively small: instead of considering only a single constraint, the repercussion-awareness test loops over all k constraints and only returns the candidate model if all k high-confidence upper bounds are at most zero. Similarly, the cost function, Algorithm 4, changes relative to Algorithm 2 in that when predicting the outcome of the repercussion-awareness test it includes this same loop over all k constraints.

H Experiments

Recall that in our experiments, we evaluate THEIA in two real-life problems: EXP-1, involving predictions about youth in the U.S. foster care system, and EXP-2, a loan-repayment setting. In both, the goal is to identify accurate models while ensuring acceptable repercussions for all demographic groups.

In EXP-1, we consider a classifier that predicts whether youth in foster care will have a job after leaving the program. Intuitively, Equation 3 captures the idea that if a youth is predicted to get a job, they are more likely to receive financial aid—leading to less severe observed repercussions. That is, R_i^ψ decreases when $\widehat{Y}_i^\psi = 1$. Notice that both the mean and variance of the repercussion resulting from a given prediction vary by race. This reflects the fact that youth from different racial backgrounds may experience systematically different outcomes, e.g., due to structural or social biases. This form aligns with the empirical findings of Chetty et al. [10], who show, using real-world data, that the distribution of financial outcomes (i.e., repercussions) differs across racial groups even when starting from similar conditions (e.g., having the same prediction $\widehat{Y}_i^\psi$).

In experiment EXP-2, we tasked THEIA with assisting banks and clients in making responsible financial decisions. In this experiment, a bank's lending decisions are informed by a classifier predicting loan repayment success. These decisions can have repercussions on clients' lives, affecting their financial well-being, savings rate, or debt-to-income ratio after a lending decision is made. In this setting, repercussions are not deterministic and can be influenced by various unobserved environmental factors. To account for this variability, our experimental setup incorporates noise models with different means and variances. This approach reflects that repercussions may differ across clients of different ages due to unobserved factors such as their professional stability or the likelihood of having a support network capable of assisting in case of economic distress. Younger individuals, e.g., who are often not yet financially stable, may experience a greater impact on their financial well-being from receiving a new loan than more financially stable senior individuals. Importantly, this setting involves complex non-linear relationships between predictions and their repercussions due to the Law of Diminishing Marginal Utility of Income: as income increases (e.g., due to an approved loan), the benefits of further increasing it decrease. This relationship is known in economics to be non-linear [28].

Furthermore, this setting presents a challenging scenario where distribution shift causes the repercussions of ML models to change between the deployment of the baseline classifier (behavior model) THEIA aims to improve upon and the time when THEIA searches for a repercussion-aware model. Specifically, it models a scenario where significant societal changes have altered the relationship between predictions and their repercussions, reflecting, for example, how loan decisions may impact clients differently over time. This allows us to empirically evaluate THEIA's robustness in cases where statistical properties of the data change after the training data was collected.

In all experiments, our implementation of THEIA used ES [36, 38] to search over the space of candidate models and the `ttest` concentration inequality. We partitioned the dataset D into D_c and D_f using a stratified sampling approach where D_c contains 60% of the data and D_f contains 40% of the data.

Experiments were conducted on a computer cluster containing 50 computer nodes with 28 cores (2 processors, 14 cores each - 56 cores with hyper-threading) Xeon E5-2680 v4 @ 2.40GHz, 128GB RAM, 200GB local SSD disk, and 50 compute nodes with 28 cores (2 processors, 18 cores each - 72 cores with hyper-threading) Xeon Gold 6240 CPU @ 2.60GHz, 192GB RAM, and 240GB local SSD disk. Each node had 3GB of allocated memory. The 500 trials of each experiment were run in parallel, and the total running time was less than 12 hours.

We now present the complete set of results for RQ1 and RQ2: does THEIA enforce repercussion constraints with high probability, while existing algorithms fail to do so; and what is the cost (e.g., in terms of accuracy) of enforcing repercussion constraints. Figure 4 shows the complete failure rate plots for both constraints across both experiments (EXP-1 and EXP-2). These experiments fully support the conclusions discussed in the main text. Specifically, they confirm that the solutions returned by THEIA *always* satisfy both repercussion constraints; other methods, by contrast, either **(1)** consistently fail to satisfy both constraints or **(2)** satisfy one but not the other.

Recall that in the main text, we presented results on EXP-1 regarding the accuracy of classifiers returned by different algorithms and the probability with which each of them returns a solution as a function of the amount of training data. In Figure 5, we present this analysis for EXP-2. In particular, Figure 5 shows, on the left, the accuracy of classifiers returned by different algorithms, and on the right, the probability of returning a solution, as a function of n, for EXP-2. Notice that THEIA's accuracy matches or exceeds that of competing methods while consistently satisfying all constraints (as shown in Figure 4). Furthermore, THEIA starts returning models that satisfy all constraints with high confidence when given a number of samples corresponding to just 0.1% of the available data. It

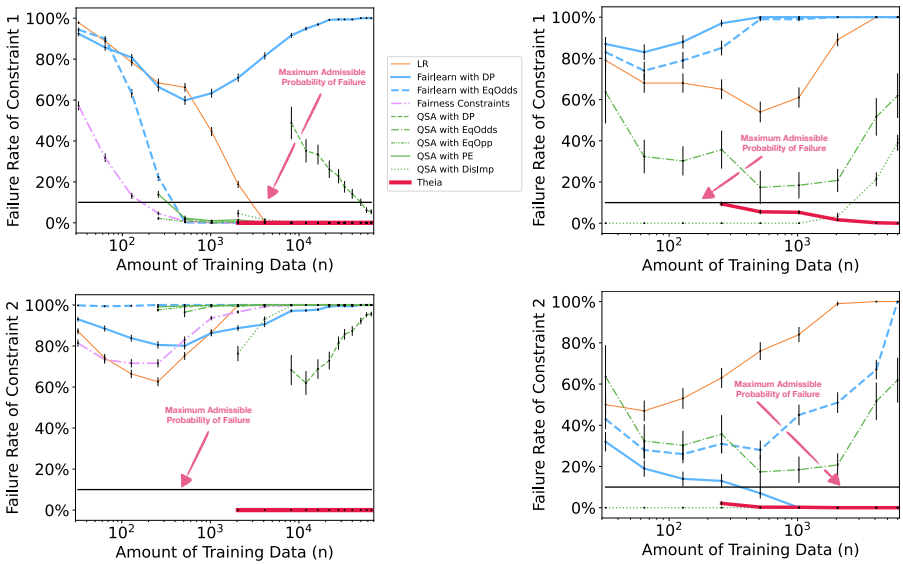


Figure 4: Failure rates of various methods w.r.t. repercussion constraints as a function of n [**Left:** EXP-1; **Right:** EXP-2]. Black lines show the maximum admissible probability of failing to enforce the constraints ($\delta_0 = \delta_1 = 10\%$).

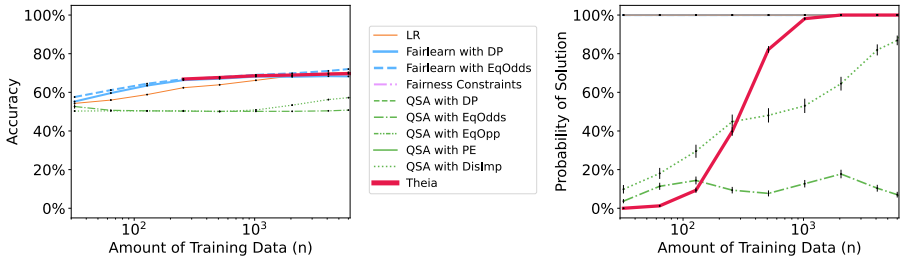


Figure 5: [**Left**: EXP-2] Accuracy of the models returned by algorithms (subject to different constraints) as a function of n. [**Right**: EXP-2] Probability that these algorithms return a solution as a function of n.

consistently returns models (with 100% probability) when given a number of samples corresponding to just 0.4% of the available data. Note that although Fairlearn and logistic regression always return a model regardless of training data size, these models never satisfy both constraints. These results demonstrate that even in a challenging setting with non-linear prediction-repercussion dependencies that are affected by distribution shift, THEIA still significantly outperforms all competing methods.

Finally, recall that in the main text we evaluated THEIA in EXP-1 under one representative value of α ($\alpha = 0.9$). We now show full results for a wide range of values of α to further support the observation that THEIA is robust with respect to various prediction-repercussion dependency levels. In particular, we investigate the performance of THEIA and competitors in terms of failure rate, probability of returning a solution, and accuracy, for different values of α and as a function of n. Notice that Figures 6 and 7 present results consistent with the observations made in Section 5; that is, the qualitative behavior of all considered algorithms remains the same for all values of α, further supporting our observation that THEIA outperforms competitors both under high prediction-repercussion dependency and when classifiers' predictions have little to no influence on the repercussions.

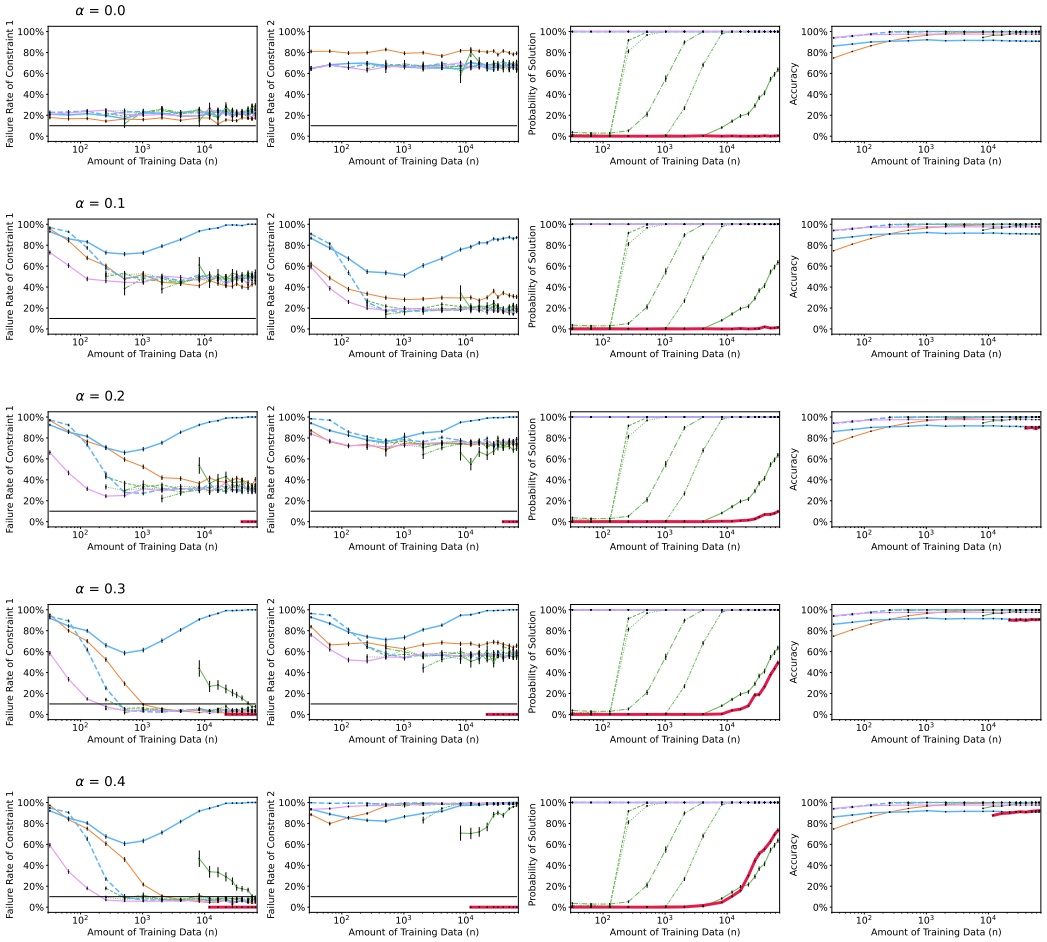


Figure 6: Algorithms' performances w.r.t. failure rate (first and second columns), probability of returning a solution (third column), and accuracy (fourth column), as a function of n and for different values of α. The black horizontal lines indicate the maximum admissible probability of failure, $\delta_0{=}\delta_1{=}10\%$. All plots use the following legend: THEIA LR QSA with DP QSA with EqOdds QSA with EqOpp QSA with PE QSA with DisImp Fairlearn with DP Fairlearn with EqOdds Fairness Constraints.

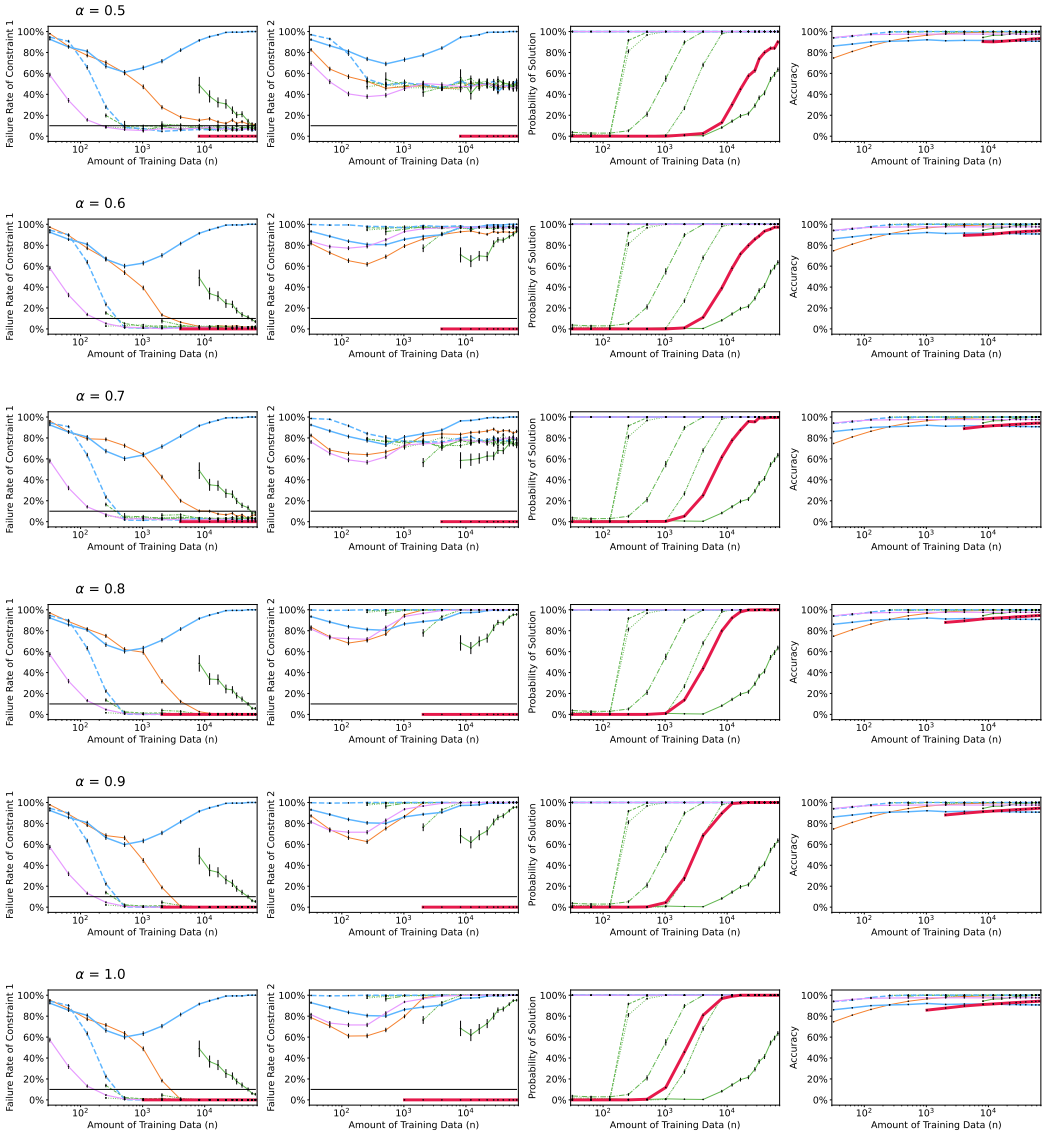


Figure 7: Algorithms' performance w.r.t. failure rate (first and second columns), probability of returning a solution (third column), and accuracy (fourth column), as a function of n and for different values of α. The black horizontal lines indicate the maximum admissible probability of failure, $\delta_0{=}\delta_1{=}10\%$. All plots use the following legend: THEIA LR QSA with DP QSA with EqOdds QSA with EqOpp QSA with PE QSA with DisImp Fairlearn with DP Fairlearn with EqOdds Fairness Constraints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are supported by theoretical analyses (Section 4) and empirical evaluation (Section 5 and Appendix H).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the keys assumptions (and when they are satisfied) required for our method to identify, with high confidence, repercussion-aware solutions. We also investigate the performance and robustness of our approach with respect to increasing levels of noise.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We present our theory using a set of four assumptions and three theorems, with complete proofs provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the experiments in detail in the main paper, and add more specific details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data for EXP-1 can be obtained directly from the NDACAN webpage. The data for EXP-2 can also be obtained online, using the reference provided. The code is not yet publicly available. We are in the final stages of developing a library that will be made publicly available to the community, and we expect to release it in the coming months. The paper contains all the necessary details for implementing the proposed method.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The main details are located in Section 5 of the main paper and Section H of the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Details can be found in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details can be found in Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not present experiments involving human subjects. Furthermore, we comply with all policies outlined by the sources from which the datasets used in EXP-1 and EXP-2 were obtained.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 4 we discuss the conditions under which the guarantees provided by our algorithm do not necessarily hold. This could lead to models that are not repercussion-aware if THEIA is deployed even when its requirements are not satisfied.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not make use of LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.