# MIT 14.662 Spring 2026, Lecture 3: Roy Models

## Part 2: Skill, learning, and choosing

David Autor, MIT and NBER

February 11, 2026 (rev 2026/02/10)

# Two applications of Roy model: Medical diagnosis and ML alignment

1. **Selection with Variation in Diagnostic Skill (Chan, Gentzkow, Yu 2022)**
   - Classification Framework
   - Structural Model
   - Identification Strategy
   - Empirical Results
   - Structural Estimation
   - Policy Implications
   - CGY Summary

2. **Misaligned by Design (Autor, Caplin, Martin, Marx 2025)**

# Selection with Variation in Diagnostic Skill

Chan, Gentzkow, Yu (*QJE* 2022)

# Motivation: Practice variation in medicine

**New puzzle** (related to Chandra/Staiger):

- Large, persistent diffs in diagnoses decisions across physicians
- E.g., some docs diagnose pneumonia much more often than others—requiring further intervention

**Main Q:** To what degree is diagnostic variation drive by skill differences vs preference differences?

**Standard interpretation:** *Variation reflects differences in preferences (risk tolerance, costs of errors)*

- Implication: This is suboptimal practice variation
- Should *standardize* decisions via uniform thresholds/guidelines

**Alternative:** *Variation may reflect differences in skill (signal quality)*

- Physicians with different skill levels (corresponding to signal quality) optimally choose different thresholds
- If decision risks are asymmetric (e.g., worse to miss a diagnosis than to over-test) and signals vary, optimal thesholds will differ
- Implication: Uniform thresholds may *reduce* welfare

**This paper:** Framework to separate *skill* from *preferences* using observational data

# Setting: Pneumonia diagnosis

**Application:** Chest X-ray interpretation for pneumonia

- Veterans Health Administration (VHA), 1999–2015
- 4.7 million cases, 3,199 radiologists, 104 stations
- Quasi-random assignment of patients to radiologists

**Key observables:**

- **Decision** $d_i \in \{0, 1\}$: Radiologist diagnoses pneumonia
- **Outcome** $m_i \in \{0, 1\}$: Patient develops pneumonia later (false negative indicator)

**Asymmetric costs:**

- If radiologist makes *false negative* diagnosis, full blown pneumonia likely
- *False positives* less costly
- Given asymmetry, doctors should err towards FP over FN

# Example chest X-rays

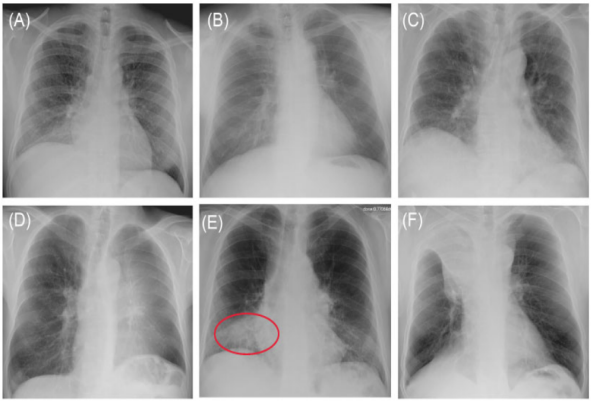



FIGURE III
Example Chest X-rays

**Training cases with expert consensus:**

(A) Miliary tuberculosis

(B) Lung nodule (cancer), left upper lobe

(C) Usual interstitial pneumonitis

(D) Left upper lobe atelectasis

(E) **Infectious pneumonia**, right lower lobe (red oval)

(F) Right upper lobe atelectasis

Only Panel E shows pneumonia—the condition radiologists must detect

Source: Fabre et al. (2018), *Diagnostic and Interventional Imaging*

# Binary classification: Setup

**State:** $s_i \in \{0, 1\}$ (patient has pneumonia or not)
**Decision:** $d_i \in \{0, 1\}$ (diagnose positive or not)

| | **Actual Positive** ($s = 1$) | **Actual Negative** ($s = 0$) |
|---|---|---|
| **Classified Positive** ($d = 1$) | True Positive (TP) | False Positive (FP) |
| **Classified Negative** ($d = 0$) | False Negative (FN) | True Negative (TN) |

**Key rates:**

$$TPR_j = \frac{TP_j}{TP_j + FN_j} \quad \text{(True Positive Rate / Sensitivity)}$$

$$FPR_j = \frac{FP_j}{FP_j + TN_j} \quad \text{(False Positive Rate / 1 - Specificity)}$$

# ‘Confusion matrix’ and ROC curve

(A) Classification Matrix

| | Actual Positive | Actual Negative |
|---|---|---|
| Classified Positive | True Positive ($TP$) | False Positive ($FP$) Type I Error |
| Classified Negative | False Negative ($FN$) Type II Error | True Negative ($TN$) |

True Positive Rate: $TPR = \frac{TP}{TP+FN}$

False Positive Rate: $FPR = \frac{FP}{FP+TN}$

(B) Receiver Operating Characteristic (ROC) Curve

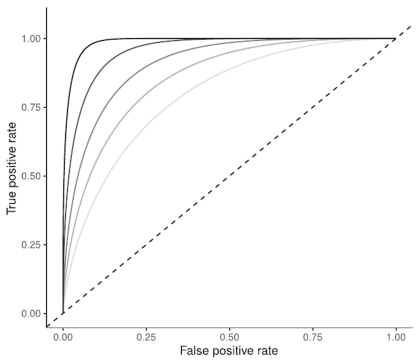



FIGURE I

Visualizing the Classification Problem

Panel A shows the standard classification matrix representing four joint outcomes depending on decisions and states. Each row represents a decision and each column represents a state. Panel B plots examples of the receiver operating characteristic (ROC) curve. It shows the relationship between the true positive rate ($TPR$) and the false positive rate ($FPR$). The particular ROC curves shown in this figure are formed assuming the signal structure in equation (5), with more accurate ROC curves (higher $\alpha_j$) further from the 45-degree line.

# ROC curve as production possibility frontier

**ROC Curve:** Plots $(FPR, TPR)$ as threshold varies

**Interpretation:**

- Frontier of achievable $(FPR, TPR)$ pairs
- **Higher skill** $\Rightarrow$ curve bows more toward $(0, 1)$
- Diagonal = random classification

**Key insight:**

- Agents on *same* ROC curve: vary in **preferences**
- Agents on *different* ROC curves: vary in **skill**

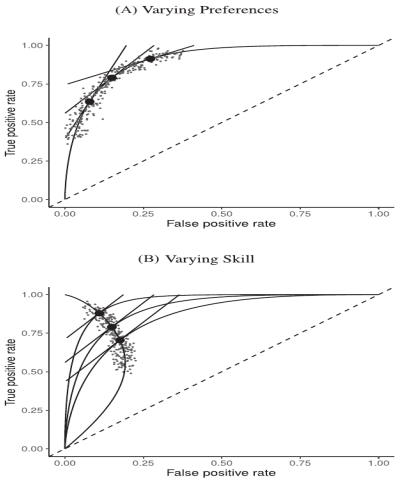



FIGURE II
Hypothetical Data Generated by Variation in Preferences versus Skill

# Signal structure

**Patient's true state:** Latent health index $\nu_i \sim \mathcal{N}(0,1)$

- Patient has pneumonia if $\nu_i > \bar{\nu}$
- $s_i = \mathbf{1}(\nu_i > \bar{\nu})$

**Radiologist's signal:** Noisy observation $w_{ij}$ for radiologist $j$

$$\begin{pmatrix} \nu_i \\ w_{ij} \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_j \\ \alpha_j & 1 \end{pmatrix} \right)$$

**Key parameter:** $\alpha_j \in (0,1]$ is radiologist $j$'s **diagnostic skill**

- $\alpha_j = 1$: Perfect signal ($w_{ij} = \nu_i$)
- $\alpha_j \to 0$: Uninformative signal
- Higher $\alpha_j \Rightarrow$ ROC curve closer to $(0,1)$ corner

# Optimal decision rule

**Radiologist's utility:**

$$u_{ij}(d, s) = \begin{cases} -1 & \text{if } d = 1, s = 0 \quad \text{(false positive)} \\ -\beta_j & \text{if } d = 0, s = 1 \quad \text{(false negative)} \\ 0 & \text{otherwise} \end{cases}$$

**Preference parameter:** $\beta_j > 0$ is relative cost of FN vs FP

- $\beta_j > 1$: Missing disease is worse than false alarm
- Expected $\beta_j > 1$ for pneumonia (serious condition)

**Optimal rule:** Threshold decision

$$d_{ij} = \mathbf{1}(w_{ij} > \tau_j^*)$$

where threshold $\tau_j^*$ depends on both skill $\alpha_j$ and preferences $\beta_j$

# Optimal threshold formula

**Physician's utility:** $\mathbb{E}[u_{ij}] = -(FP_j + \beta_j \cdot FN_j)$

**Key result:** The optimal diagnostic threshold is:

$$\tau^*(\alpha_j, \beta_j) = \frac{\bar{\nu} - \sqrt{1-\alpha_j^2} \cdot \Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)}{\alpha_j}$$

where $\Phi^{-1}$ is the inverse standard normal CDF

**Comparative statics: When to call pneumonia**

- Higher $\bar{\nu}$ (lower prevalence) $\Rightarrow$ higher threshold
- Higher $\beta_j$ (FN more costly) $\Rightarrow$ lower threshold (diagnose more)
- Effect of $\alpha_j$ (diagnostic skill) is **ambiguous**:
  - Shifts posterior mean toward $\bar{\nu}$ (increases threshold)
  - Reduces posterior variance (decreases threshold)

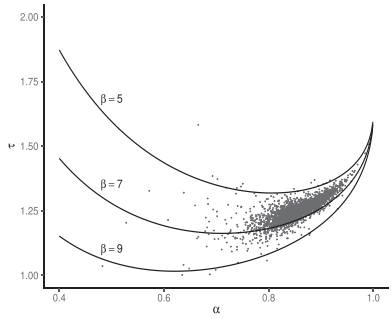



FIGURE IX

Optimal Diagnostic Threshold

This figure shows how the optimal diagnostic threshold varies as a function of skill $\alpha$ and preferences $\beta$ with iso-preference curves for $\beta \in \{5, 7, 9\}$. Each iso-preference curve illustrates how the optimal diagnostic threshold varies with the evaluation skill for a fixed preference, given by equation (7), using $\bar{\nu} = 1.635$ estimated from the model. Dots on the figure represent the empirical Bayes posterior mean of $\alpha$ (on the $x$-axis) and $\tau$ (on the $y$-axis) for each radiologist. The empirical Bayes posterior means are the same as those shown in Online Appendix Figure A.13. Details on the empirical Bayes procedure are given in Online Appendix E.3.

# Identification challenge

**Problem:** We observe $(d_i, m_i)$ but not true state $s_i$

- $m_i = 1$ only if $d_i = 0$ and $s_i = 1$ (false negative revealed)
- Cannot observe false positives directly

**Observable moments for each radiologist:**

$$P_j = \Pr(d_i = 1|j) = TP_j + FP_j \quad \text{(diagnosis rate)}$$
$$FN_j = \Pr(m_i = 1|j) \quad \text{(miss rate)}$$

**Notice:** Given share at risk $S = TP_j + FN_j$, can solve for full classification matrix:

$$TP_j = S - FN_j$$
$$FP_j = P_j - TP_j$$
$$TN_j = 1 - S - FP_j$$

# Quasi-random assignment

**Assumption 1 (Conditional Independence):**

- Conditional on hour, day of week, month, location, and patient state $s_i$, potential diagnosis decisions $\{d_{ij}\}_{j \in \mathcal{J}}$ are independent of assigned radiologist $\hat{j}(i)$

**Institutional basis:**

- Radiologists assigned based on who is "on call"
- Assignment determined by scheduling, not patient characteristics
- Control for time-station fixed effects $\mathbf{T}_i$

**Implication:** Can use radiologist assignment as instrument for diagnosis

- Leave-out diagnosis propensity $Z_i$: predicted diagnosis using other radiologists' patients
- Leave-out miss rate $\hat{m}_i$: predicted miss rate

# Testing for skill heterogeneity

**Key prediction:** If skill is uniform, radiologists lie on *same* ROC curve

**Monotonicity condition:** Under uniform skill,

$$\text{Higher } P_j \Rightarrow \text{Lower } FN_j$$

(diagnosing more $\Rightarrow$ missing fewer cases)

**Test:** Regress miss rate on diagnosis rate using IV

$$m_i = \Delta \cdot d_i + \mathbf{X}_i'\gamma + \varepsilon_i$$

Instrument: $Z_i$ (leave-out diagnosis propensity)

**Under uniform skill:** $\Delta < 0$ (monotonicity holds)

**With skill heterogeneity:** $\Delta$ can be positive

- High-diagnosing radiologists may have *worse* signals
- Both diagnose more AND miss more cases

# Main result: Positive slope

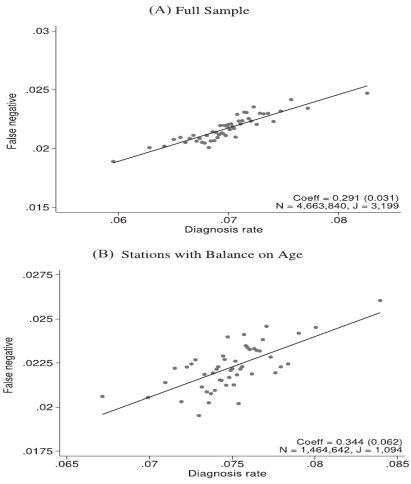



FIGURE VI

Diagnosis and Miss Rates

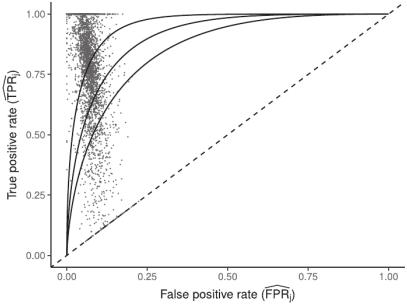



FIGURE V

Projecting Data on ROC Space

This figure plots the true positive rate ($\widehat{TPR}_j$) and false positive rate ($\widehat{FPR}_j$) for each radiologist across the 3,199 radiologists in our sample who have at least 100 chest X-rays. The figure is based on observed risk-adjusted diagnosis and miss rates $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN}_j^{\text{obs}}$, then adjusted for the share of X-rays not at risk for pneumonia ($\hat{\kappa} = 0.336$) and the share of cases in which pneumonia first manifests after the initial visit ($\hat{\lambda} = 0.026$). The values of $\widehat{TPR}_j$ and $\widehat{FPR}_j$ are then computed using the estimated prevalence rate $\hat{S} = 0.051$. Values are truncated to impose $\widehat{TPR}_j \leqslant 1$ (affects 597 observations), $\widehat{FPR}_j \geqslant 0$ (affects 44 observations), and $\widehat{TPR}_j \geqslant \widehat{FPR}_j$ (affects 68 observations). See Section IV.C and Online Appendix C for more details.

# Interpreting the positive slope

**Finding:** $\hat{\Delta} > 0$ — radiologists who diagnose more also miss more

**Implication: Monotonicity is violated**

- Cannot be explained by preference differences alone
- Requires heterogeneity in diagnostic skill

**Intuition:**

- Low-skill radiologists have noisier signals
- To achieve acceptable miss rate, must diagnose more patients
- Despite diagnosing more, still miss more cases

**Magnitude:**

- Full sample: $\hat{\Delta} = 0.291$ (s.e. $= 0.031$)
- Balanced on age: $\hat{\Delta} = 0.344$ (s.e. $= 0.082$)

# Estimation approach (big picture)

**Data:** Diagnoses $d_i$ and revealed false negatives $m_i$

- Key constraint: $m_i$ only observed when $d_i = 0$
- Cannot directly observe false positives

**Identification:** Cross-radiologist variation in $(P_j, FN_j)$ identifies $(\alpha_j, \beta_j)$

- Each radiologist characterized by sufficient statistics: $(n_j^d, n_j^m, n_j)$
- Model maps skill/preferences to predicted diagnosis and miss rates

**Hierarchical structure:**

- $(\alpha_j, \beta_j)$ drawn from joint distribution with hyperparameters $\boldsymbol{\theta}$
- Allows for correlation between skill and preferences
- Estimated via maximum likelihood

# Estimation approach (detailed)

**Data:** Observe diagnoses $d_i$ and false negatives $m_i$ (note: $m_i = 0$ if $d_i = 1$)

**Define probabilities** conditional on $\gamma_j \equiv (\alpha_j, \beta_j)$:

$$
\begin{aligned}
p_{1j}(\gamma_j) &\equiv \Pr(w_{ij} > \tau_j^* \,|\, \gamma_j) \quad \text{(positive diagnosis)} \\
p_{2j}(\gamma_j) &\equiv \Pr(w_{ij} < \tau_j^*, \nu_i > \bar{\nu} \,|\, \gamma_j) \quad \text{(revealed FN)} \\
p_{3j}(\gamma_j) &\equiv \Pr(w_{ij} < \tau_j^*, \nu_i < \bar{\nu} \,|\, \gamma_j) \quad \text{(true negative)}
\end{aligned}
$$

**Likelihood for case $i$:**

$$
\mathcal{L}_i(d_i, m_i | \gamma_{j(i)}) = \begin{cases} (1-\kappa) p_{1j}, & \text{if } d_i = 1 \\ (1-\kappa)(p_{2j} + \lambda p_{3j}), & \text{if } d_i = 0, m_i = 1 \\ (1-\kappa)(1-\lambda) p_{3j} + \kappa, & \text{if } d_i = 0, m_i = 0 \end{cases}
$$

where $\kappa$ = share never at risk, $\lambda$ = prob. FN among not-at-risk is revealed

# Estimation approach (cont.)

**Sufficient statistics** for radiologist $j$'s likelihood:

- $n_j^d = \sum_{i \in I_j} d_i$ (number of positive diagnoses)
- $n_j^m = \sum_{i \in I_j} m_i$ (number of revealed false negatives)
- $n_j = |I_j|$ (total cases assigned to $j$)

**Hierarchical structure:**

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim f(\alpha, \beta | \boldsymbol{\theta})$$

**Parametrization:**

- $\alpha_j = \frac{1}{2}(1 + \tanh \tilde{\alpha}_j) \in (0, 1)$
- $\beta_j = \exp \tilde{\beta}_j > 0$
- $(\tilde{\alpha}_j, \tilde{\beta}_j)$ jointly normal with hyperparameters $\boldsymbol{\theta}$

# Structural estimates

TABLE I

STRUCTURAL ESTIMATION RESULTS

Panel A: Model parameter estimates

| | Estimate | Description |
|---|---|---|
| $\mu_\alpha$ | 0.945 | Mean of $\tilde{\alpha}_j$, $\alpha_j = \frac{1}{2}(1 + \tanh \tilde{\alpha}_j)$ |
| | (0.219) | |
| $\sigma_\alpha$ | 0.296 | Standard deviation of $\tilde{\alpha}_j$ |
| | (0.029) | |
| $\mu_\beta$ | 1.895 | Mean of $\tilde{\beta}_j$, $\beta_j = \exp \tilde{\beta}_j$ |
| | (0.249) | |
| $\sigma_\beta$ | 0.136 | Standard deviation of $\tilde{\beta}_j$ |
| | (0.044) | |
| $\lambda$ | 0.026 | Share of at-risk negatives developing subsequent |
| | (0.001) | pneumonia |
| $\bar{\nu}$ | 1.635 | Prevalence $S = 1 - \Phi(\bar{\nu})$ |
| | (0.091) | |
| $\kappa$ | 0.336 | Share not at risk for pneumonia |

Panel B: Radiologist posterior means

| | | Percentiles | | | |
|---|---|---|---|---|---|
| | Mean | 10th | 25th | 75th | 90th |
| $\alpha$ | 0.855 | 0.756 | 0.816 | 0.908 | 0.934 |
| | (0.050) | (0.079) | (0.065) | (0.035) | (0.025) |
| $\beta$ | 6.713 | 5.596 | 6.071 | 7.284 | 7.909 |
| | (1.694) | (1.608) | (1.659) | (1.750) | (1.780) |
| $\tau$ | 1.252 | 1.165 | 1.208 | 1.298 | 1.336 |
| | (0.006) | (0.009) | (0.006) | (0.008) | (0.012) |

*Notes.* This table shows model parameter estimates (Panel A) and moments in the implied distribution of empirical Bayes posterior means across radiologists (Panel B). $\mu_\alpha$ and $\sigma_\alpha$ determine the distribution of radiologist diagnostic skill $\alpha$, and $\mu_\beta$ and $\sigma_\beta$ determine the distribution of radiologist preferences $\beta$ (the disutility of a false negative relative to a false positive). We assume that $\alpha$ and $\beta$ are uncorrelated. $\lambda$ is the proportion of at-risk chest X-rays with no radiographic pneumonia at the time of exam but subsequent development of pneumonia. $\bar{\nu}$ describes the prevalence of pneumonia at the time of the exam among at-risk chest X-rays. $\kappa$ is the proportion of chest X-rays not at risk for pneumonia. It is calibrated as the proportion of patients with predicted probability of pneumonia less than 0.01 from a random forest model of pneumonia based on rich characteristics in the patient chart. Parameters are described in further detail in Sections V.A and V.B. The method to calculate empirical Bayes posterior means is described in Online Appendix E.3. Standard errors, shown in parentheses, are computed by block bootstrap, with replacement, at the radiologist level.

# Key structural findings

**Skill ($\alpha$):**

- Mean skill: $\bar{\alpha} = 0.85$ (correlation between signal and true state)
- 10th percentile: 0.76, 90th percentile: 0.93
- Substantial heterogeneity in diagnostic ability

**Preferences ($\beta$):**

- Mean: $\bar{\beta} = 6.71$ (FN costs 6.7$\times$ more than FP)
- 10th percentile: 5.60, 90th percentile: 7.91
- Less heterogeneity than in skill

**Decomposition:**

- **39%** of variation in *decisions* from skill heterogeneity
- **78%** of variation in *outcomes* (miss rates) from skill

# Welfare framework

**Over to you, Amy!**

**Social planner's welfare index:**

$$W = 1 - \frac{FP + \beta^s \cdot FN}{FP^0 + \beta^s \cdot FN^0}$$

where $\beta^s$ is social cost of FN relative to FP

**Status quo:** $W = 0$
**First best:** $W = 1$

**Counterfactual policies:**

1. **Fixed threshold:** Impose uniform $\tau$ across all radiologists
2. **Optimal threshold:** Allow $\tau^*(\alpha_j, \beta^s; \vec{\nu})$ to vary with skill
3. **Improve skill:** Raise all $\alpha_j$ to higher percentile

**Wait, this is a labor class...**

# Summary: Chan, Gentzkow, Yu

**Contribution: A Roy model of decision-making with skill**

- Framework to separate skill vs. preferences in diagnostic decisions
- Use ROC curve as production possibility frontier
- Identify skill heterogeneity from violations of monotonicity

**Empirical findings:**

- Positive correlation between diagnosis and miss rates $\Rightarrow$ skill varies
- 39% of decision variation, 78% of outcome variation from skill
- Average skill $\alpha = 0.85$, average preference $\beta = 6.71$

**Economic implications:**

- Generalizes to other settings: judges, teachers, loan officers

# Misaligned by Design: Incentive Failures in Machine Learning

Autor, Caplin, Martin, Marx (2025)

# Motivation

**This paper is closely related to Chan-Gentzkow-Yu**

- CGY distinguish: Choosing different points on PPF/ROC (preference/risk) **vs** having different PPFs/ROC curves (diagnostic skill).
- ACMM show how ML design choices can distort learning about PPF/ROC by incentivizing model to choose the right point on the PPF/ROC

# Context: ML and decision-making

- AI is increasingly deployed in high-stakes domains (medicine, finance, etc.).
- Error costs in these domains are often **asymmetric**: false positives vs. false negatives.
  - Misdiagnosing pneumonia when absent is an inconvenience, but failing to detect it when present can be life-threatening (as per Chan, Gentzkow, and Yu).
- Critical that AI systems are **'aligned'** with the intentions of their human deployers (Bostrom 2014; Bengio et al. 2023; Ji et al. 2023).
- Standard practice to improve alignment: account for asymmetric costs when training AI.

# Alignment Premise

## Aligned Learning Premise (ALP)

Training on human objectives $\rightarrow$ better alignment with those objectives.

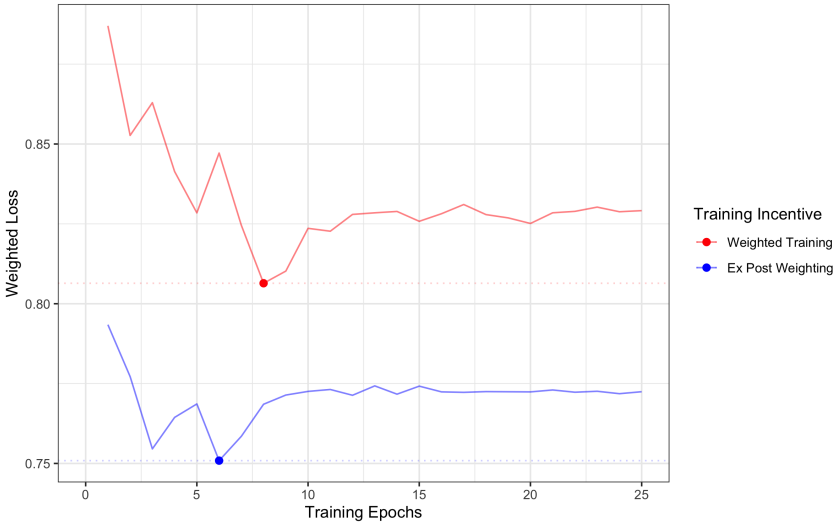
- Common implementation:
  - Weighted loss functions
  - Weighted resampling
- Paper finds that ALP is false in two focal applications. (Will show only one here)
- Better results when:
  1. Train without accounting for human objectives.
  2. Apply human objectives ex post to predictions.

# Focal Applications

1. Pneumonia diagnosis with chest X-rays (architecture: deep neural networks).
   - Widely-adopted in the field, this algorithm has also been used to study joint human and AI decision-making.
   - A variant leveraged by Agarwal et al. (2023) to study how expert radiologists make decisions when aided by AI recommendations.
2. Image classification with CIFAR (architecture: transformers).
   - A leading benchmark task for visual prediction (Krizhevsky and Hinton 2009).
   - Same family of architectures used to train LLMs.

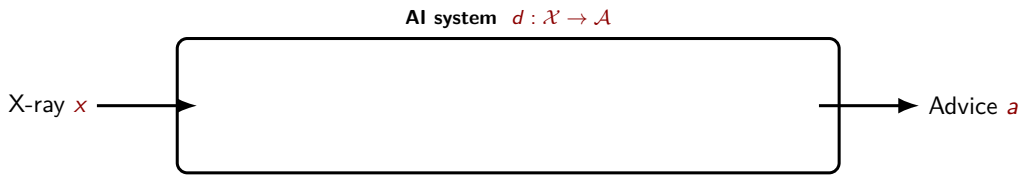
For both: unweighted training + ex-post adjustment (*Ex-Post Weighting*) outperforms utility-weighted training (*Weighted Training*).

# Empirical Result: Pneumonia



# Theoretical Framework

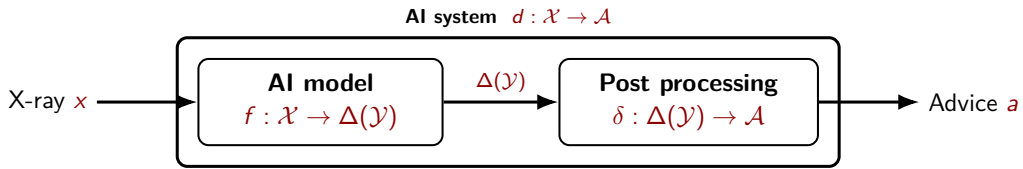
- To understand this result, create a theoretical framework of alignment.
- Imagine that a machine learning engineer (ME) needs to develop an AI system to provide advice ($a \in \mathcal{A}$) based on an X-ray ($x \in \mathcal{X}$) with a disease or not ($y \in \mathcal{Y}$).
- The AI system is summarized as a function $d : \mathcal{X} \to \mathcal{A}$ from X-rays to advice.




- The ME has utility $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}_+$ over advice conditional on the disease state, which can encode the preferences of downstream users for advice.
- Growing literature considers how a doctor or hospital might use such advice to generate diagnosis decisions.

# Theoretical Framework

- The ME trains an AI model to be the key element in this AI system.




- The AI model outputs a probability of each disease state $y$ based on any X-ray $x$ and is summarized by a prediction function $f : \mathcal{X} \to \Delta(\mathcal{Y})$.
- To generate advice, the ME adds post processing $\delta : \Delta(\mathcal{Y}) \to \mathcal{A}$ (re-calibration, thresholding, etc.) that maps the output of the AI model to specific advice $a$.
- **Alignment** is achieved if the AI model the ME trains, in combination with optimal post processing, yields an AI system that maximizes the ME's expected utility.

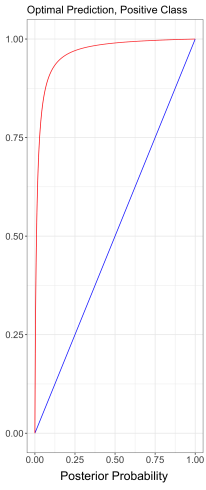
# Theoretical Framework

- In practice, the ME has many ways to try to achieve alignment, but we focus on a common one, which is the choice of loss function $\ell$ for training the AI model.
- Our economic approach considers the choice of loss function $\ell$ to achieve alignment as an incentive design problem for a downstream agent, which is the machine learner.
- Based on the incentives provided by the loss function, the machine learner's problem is to find a prediction model $f : \mathcal{X} \to \Delta(\mathcal{Y})$ to minimize expected loss $\mathbb{E}[\ell(f(X), Y)]$.
- We model the machine learner as one would model a human learner in economics: as first selecting an information structure ($\Delta(\Delta(\mathcal{Y}))$) and then choosing a "prediction" ($p \in \Delta(\mathcal{Y})$) based on the realized distribution, both to minimize expected losses.
- Frictions prevent machine learner from selecting perfectly informative structure, so incentives matter.

# Why Does ALP Fail?

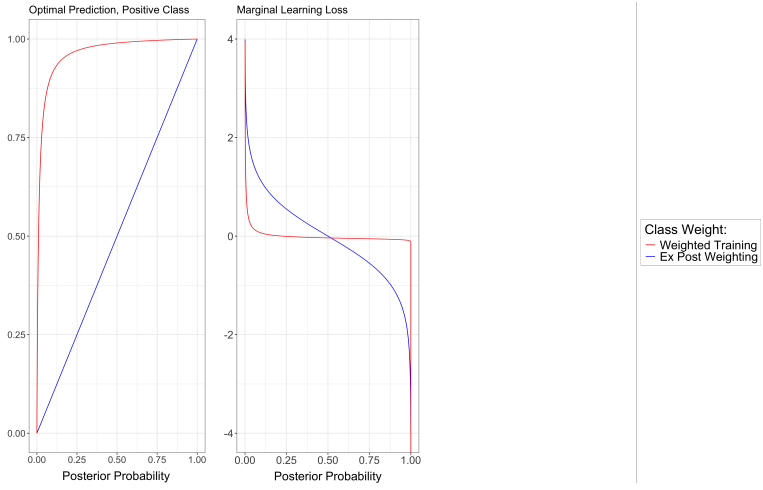
- **Key insight: Machine learners face two incentivized tasks**
  1. Choosing: the prediction to make given a probability over disease states.
  2. Learning: selecting an information structure (a probability distribution over probability distributions over disease states).
- Intuition might suggest that learning is not an incentive problem: the machine should simply learn as effectively as possible.
- But the mathematics of machine learning dictate otherwise: loss function dictates the shape of the gradient.
- Asymmetric/weighted loss $\Rightarrow$ aligns (1) but distorts (2) $\Rightarrow$ misalignment.

# Incentives to Learn



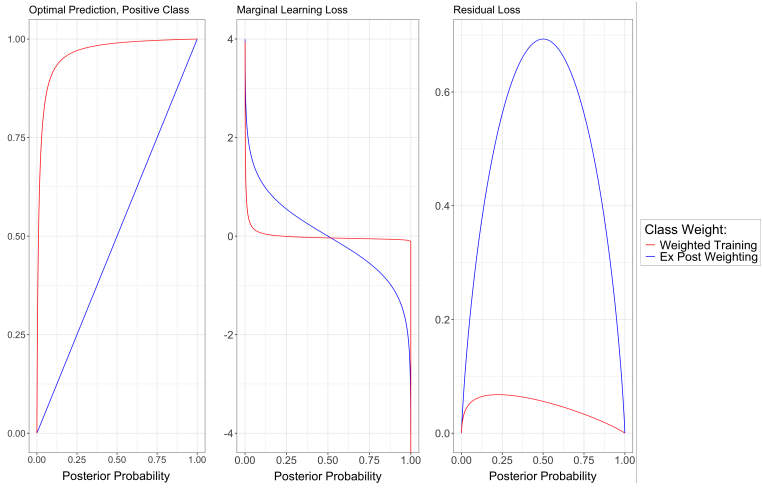

Weighted loss inflates predictions $\Rightarrow$

# Incentives to Learn




Weighted loss inflates predictions $\Rightarrow$ dampens marginal incentives to learn.

# Incentives to Learn




Weighted loss inflates predictions $\Rightarrow$ dampens incentives to learn.

# Broader Implications

- Popular methods (weighted losses, resampling) are intuitive but misleading.
- Gains on benchmarks may mask misaligned incentives.
- Need theoretical foundation, not just an engineering approach.

# Conclusion: ACMM

- Machine learning is an incentive problem.
- Distinguish
  1. Incentives for choice.
  2. Incentives for learning.
- Misalignment arises when incentives for learning are distorted
  - Popular methods (weighted losses, resampling) are intuitive but misleading.
  - Gains on benchmarks may mask misaligned incentives.
  - (Methodological point: Need theoretical foundation, not just an engineering approach.)

# Final Takeaway

- Alignment = objectives + incentives to learn.
- Misaligned by design $\rightarrow$ predictable failures.
- Economics offers a missing lens

# References


- Agarwal, Nikhil et al. (2023). *Combining human expertise with artificial intelligence: Experimental evidence from radiology*. Tech. rep. National Bureau of Economic Research.
- Bengio, Yoshua et al. (2023). "Managing AI risks in an era of rapid progress". In: *arXiv preprint arXiv:2310.17688*.
- Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Ji, Jiaming et al. (2023). "AI alignment: A comprehensive survey". In: *arXiv preprint arXiv:2310.19852*.
- Krizhevsky, Alex and Geoffrey Hinton (2009). "Learning multiple layers of features from tiny images". Technical Report, University of Toronto.