Where the Human Goes: Optimal Allocation of Expert Review in Machine-Assisted Financial Diligence

Daniel Ray Edgar
Finsider Labs

October 2026

Abstract

"Human in the loop" is a slogan until three questions have exact answers: which machine outputs a human must see, how many of the unseen outputs must be sampled before the rest can be certified, and what review can and cannot remove. I show that all three have closed-form answers under stated assumptions. First, when reviewing item i costs c_i and the item carries dollar exposure a_i, expected loss is minimised by reviewing exactly the items whose expected dollar error $a_i(1-q_i)$ exceeds c_i, where q_i is the machine's posterior probability of being right. The confidence threshold is therefore item-specific, $1-c_i/a_i$, and a single global cut-off is optimal only when cost is proportional to exposure. Under a review budget the problem is a knapsack, and a density-ordered greedy rule is within a factor of two of optimal. Second, the unreviewed remainder can be certified without trusting the machine's confidence: a blind uniform sample of $n \geq \ln(1/\delta)/\ln(1/(1-\varepsilon))$ items with zero errors bounds the error rate by ε at confidence $1-\delta$, and sampling in proportion to exposure bounds the fraction of dollars in error. Third, a reviewer who sees the machine's answer and accepts a wrong one with probability β leaves a residual error of at least β times the machine's error, a floor that a second anchored reviewer lowers by less than independence would suggest. I compose the results into a release gate with an explicit guarantee and state what they do not establish.

1 Introduction

Financial diligence in a merger or acquisition consists of a very large number of small judgments: this transaction is a one-time expense, this receipt matches that invoice, this balance is classified correctly. Software now makes most of these judgments, and makes most of them correctly. The practical question is what to do about the rest. The standard answer, a "human in the loop", does not say which items the human looks at, how the unreviewed items acquire any assurance, or how much error survives a review.

The human-factors literature gives reasons to take these questions seriously. Parasuraman and Riley [1997] distinguish appropriate use of automation from misuse (over-reliance) and disuse (under-reliance). Parasuraman and Manzey [2010] review the evidence that operators monitoring a mostly reliable automated system become complacent and exhibit automation

bias, accepting automated output in place of their own checking. Tversky and Kahneman [1974] describe anchoring: estimates that start from a supplied value are adjusted away from it insufficiently. On the machine side, Guo et al. [2017] show that the confidence scores of modern neural networks are often miscalibrated, so a stated confidence of 0.99 cannot simply be read as a one per cent error rate.

This paper turns the slogan into three precise statements. *Allocation* (Sections 3 and 4): reviewing item i is optimal exactly when its expected dollar error exceeds its review cost. This extends the reject rule of Chow [1970] from a uniform error cost to item-specific dollar stakes, and it implies that the usual single confidence cut-off is the wrong shape; under a hard budget the allocation is a knapsack with a greedy solution of guaranteed quality. *Certification* (Sections 5 and 6): a blind random sample of the released items yields an exact binomial bound on their error rate [Clopper and Pearson, 1934], with a closed-form sample size, whether or not the machine's confidence is meaningful. *The anchoring floor* (Section 7): review that starts from the machine's answer cannot push residual error below a fixed multiple of the machine's error.

Section 8 treats reviewer quality without ground truth, Section 9 composes the pieces into a release gate, and Section 10 works an illustrative example. The ingredients are classical; what I add is the dollar-weighted rule with the proof that a global threshold fails, the anchoring floor and its dependence corollary, the corrected sample size under imperfect review of the sample, and the composition into one gate. The paper contains no empirical results.

2 Setup and Notation

There are N items, each a single machine decision whose correctness matters in dollars: a classification, a match, an extracted figure.

Definition 2.1 (Items)**.** Item i has *exposure* $a_i > 0$, the dollar loss incurred if the machine's decision on i is wrong and is released uncorrected. The machine reports a *confidence* $q_i \in [0, 1]$, and I write $r_i = 1 - q_i$. A human review of item i costs $c_i > 0$ dollars. Let W_i denote the event that the machine's decision on item i is wrong.

Two assumptions are used in the allocation results and only there. **(A1), meaningful confidence:** q_i is the posterior probability, given everything the allocation rule may see, that the machine's decision on item i is correct, so $\mathbb{P}(W_i) = r_i$. **(A2), review resolves the item:** a reviewed item is released correct, so its loss after review is zero.

Assumption (A1) is strong [Guo et al., 2017]; the certification results of Sections 5 and 6 do not use it. Assumption (A2) is relaxed in Section 7. A *review policy* is a set $S \subseteq \{1, \dots, N\}$ of items sent to review, chosen as a function of the observable triples (a_i, q_i, c_i). Its complement $U = \{1, \dots, N\} \setminus S$ is the *released set*. Under (A1) and (A2) the expected total cost of policy S, review cost plus expected error loss, is

$$L(S) \;=\; \sum_{i \in S} c_i \;+\; \sum_{i \notin S} a_i r_i. \tag{1}$$

No independence between items is needed for (1); it follows from linearity of expectation. I call $v_i = a_i r_i$ the *expected dollar error* of item i.

3 The Dollar-Weighted Reject Rule

Chow [1970] studied a recogniser that may decline to decide and showed that, when every error costs the same and every rejection costs the same, the optimal rule rejects exactly the patterns whose maximum posterior probability falls below a fixed threshold. In diligence the errors do not cost the same: a misclassified USD 500,000 item and a misclassified USD 900 item are different events.

Theorem 3.1 (Dollar-weighted reject rule)**.** *Assume* (A1) *and* (A2)*. The expected total cost* (1) *is minimised over all review policies by*

$$S^\star \;=\; \{\, i : a_i r_i > c_i \,\},$$

and $L(S^\star) = \sum_{i=1}^{N} \min(c_i,\, a_i r_i)$*. A policy* S *is optimal if and only if it contains every item with* $a_i r_i > c_i$ *and no item with* $a_i r_i < c_i$*.*

Proof. For any policy S, write the cost as a sum over items: $L(S) = \sum_{i=1}^{N} \ell_i(S)$ with $\ell_i(S) = c_i$ if $i \in S$ and $\ell_i(S) = a_i r_i$ otherwise. Each term satisfies $\ell_i(S) \geq \min(c_i, a_i r_i)$, so $L(S) \geq \sum_i \min(c_i, a_i r_i)$. The policy $S^\star$ attains the minimum in every term simultaneously, because membership of item i in S places no constraint on any other item. Hence $S^\star$ is optimal. Finally, $L(S) = L(S^\star)$ holds if and only if $\ell_i(S) = \min(c_i, a_i r_i)$ for every i, which fails exactly when S omits an item with $a_i r_i > c_i$ or includes an item with $a_i r_i < c_i$. □

Corollary 3.2 (Item-specific confidence threshold)**.** *Under the assumptions of Theorem 3.1, item* i *is reviewed by* $S^\star$ *if and only if* $q_i < \tau_i := 1 - c_i/a_i$*.*

Proof. $a_i r_i > c_i$ is equivalent to $1 - q_i > c_i/a_i$. □

When review cost is roughly constant, τ_i rises towards one as exposure grows: a large item is reviewed at a confidence at which a small item is released (Figure 1; Section 10 quantifies one case). The common practice of one global cut-off ("review everything below 0.90") is the policy G_τ: for $\tau \in \mathbb{R}$, let $G_\tau = \{\, i : q_i < \tau \,\}$.

Theorem 3.3 (A global threshold is optimal only for proportional costs)**.** *Fix exposures and costs with* $c_i < a_i$ *for every* i*, and write* $\rho_i = c_i/a_i$*.*

(a) *If* $\rho_i = \rho$ *for all* i*, then* $G_{1-\rho}$ *is optimal for every confidence profile* $(q_1, \ldots, q_N)$*.*

(b) *If* $\rho_j < \rho_k$ *for some pair* j, k*, then for every* $\tau \in \mathbb{R}$ *there is a confidence profile on which* $L(G_\tau) > L(S^\star)$*.*

Proof. (a) By Corollary 3.2, $S^\star = \{i : q_i < 1 - \rho\} = G_{1-\rho}$.

(b) By the pointwise argument in the proof of Theorem 3.1, every item contributes at least as much to $L(G_\tau)$ as to $L(S^\star)$, so it suffices to exhibit one item whose contribution is strictly larger. Note $0 < 1 - \rho_k < 1 - \rho_j < 1$. If $\tau > 1 - \rho_k$, choose q_k with $1 - \rho_k < q_k < \min(\tau, 1)$. Then G_τ reviews item k at cost c_k, while $a_k r_k = a_k(1 - q_k) < a_k \rho_k = c_k$. If instead $\tau \leq 1 - \rho_k$, choose q_j with $\max(\tau, 0) < q_j < 1 - \rho_j$, a non-empty interval. Then $q_j \geq \tau$, so G_τ releases item j at expected loss $a_j(1 - q_j) > a_j \rho_j = c_j$. In either case the remaining confidences may be chosen arbitrarily. □

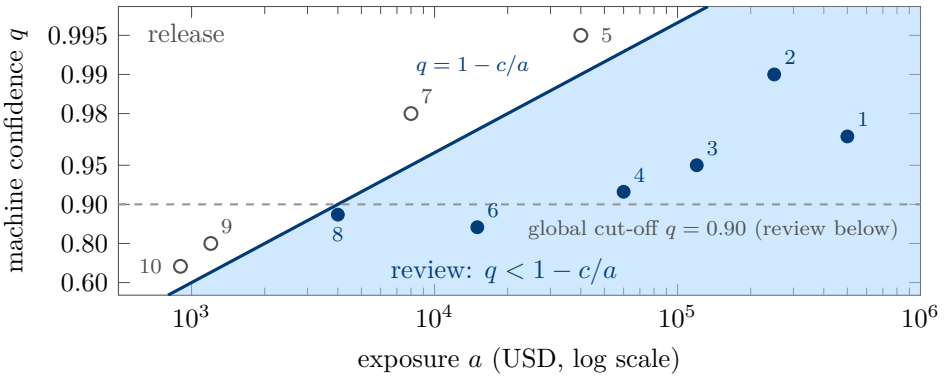


Figure 1: The review region for review cost $c =$ USD 400. Confidence q is drawn on a logarithmic scale in $1 - q$, so the optimal boundary $q = 1 - c/a$ of Corollary 3.2 is a straight line. The shaded region is reviewed under the dollar-weighted rule; the dashed line is a global cut-off at 0.90, which reviews only what lies below it. Points are the items of Table 1 (filled: reviewed under the dollar-weighted rule; hollow: released).

4 Review under a Budget

Theorem 3.1 prices review in the same dollars as error. Often review capacity is instead fixed in advance at B dollars (or hours), and the objective is to remove as much expected dollar error as the budget allows:

$$\max_{S \subseteq \{1,\dots,N\}} \sum_{i \in S} v_i \quad \text{subject to} \quad \sum_{i \in S} c_i \le B, \qquad v_i = a_i r_i. \tag{2}$$

This is the 0-1 knapsack problem, which is hard to solve exactly in general. Its linear relaxation, in which an item may be reviewed fractionally, was solved by Dantzig [1957]: fill in decreasing order of value per unit cost. I reprove that fact because the bound that follows depends on it.

Throughout this section, discard any item with $c_i > B$ (it cannot be reviewed), assume $\sum_i c_i > B$ (otherwise review everything), and order the items by *density* $d_i = v_i / c_i$ so that $d_1 \ge d_2 \ge \cdots \ge d_N$. Let m be the largest index with $\sum_{i \le m} c_i \le B$; then $1 \le m < N$, and item $m + 1$ is the *break item*. The *greedy prefix* is $\{1, \dots, m\}$ with value $G = \sum_{i \le m} v_i$.

Lemma 4.1 (Fractional optimum)**.** *The linear programme* $\max \sum_i v_i x_i$ *subject to* $\sum_i c_i x_i \le B$ *and* $0 \le x_i \le 1$ *has optimal value*

$$F \;=\; \sum_{i \le m} v_i \;+\; \frac{B - \sum_{i \le m} c_i}{c_{m+1}}\, v_{m+1},$$

attained by $x_i = 1$ *for* $i \le m$, $x_{m+1} = (B - \sum_{i \le m} c_i)/c_{m+1}$ *and* $x_i = 0$ *for* $i > m + 1$.

Proof. Let $d = d_{m+1}$. For any feasible x,

$$\sum_i v_i x_i \;=\; \sum_i (v_i - d c_i)\, x_i \;+\; d \sum_i c_i x_i \;\le\; \sum_{i \le m} (v_i - d c_i) \;+\; dB,$$

because $v_i - dc_i = c_i(d_i - d)$ is non-negative for $i \le m$ and non-positive for $i > m$, each x_i lies in $[0,1]$, $d \ge 0$ and $\sum_i c_i x_i \le B$. The stated x is feasible, since $0 \le x_{m+1} < 1$ by the definition of m and its total cost is exactly B. Its value is $\sum_{i \le m} v_i + x_{m+1} v_{m+1}$, and since $v_{m+1} - dc_{m+1} = 0$ this equals $\sum_{i \le m}(v_i - dc_i) + d\,(\sum_{i \le m} c_i + x_{m+1} c_{m+1}) = \sum_{i \le m}(v_i - dc_i) + dB$, the upper bound. □

Theorem 4.2 (Half-optimality of greedy with a fallback)**.** *Let* OPT *be the optimal value of* (2) *and let* $v_{\max} = \max_i v_i$ *be the largest single expected dollar error among items that fit the budget. Then*

$$\max(G,\, v_{\max}) \;\ge\; \tfrac{1}{2}\,\mathrm{OPT} \qquad \textit{and} \qquad G \;\le\; \mathrm{OPT} \;\le\; F \;\le\; G + v_{m+1}.$$

Proof. Every feasible set for (2) is feasible for the linear programme of Lemma 4.1, so $\mathrm{OPT} \le F$. The greedy prefix is feasible, so $G \le \mathrm{OPT}$. By Lemma 4.1, $F = G + x_{m+1} v_{m+1}$ with $0 \le x_{m+1} < 1$, hence $F \le G + v_{m+1}$. Therefore $\mathrm{OPT} \le G + v_{m+1} \le G + v_{\max} \le 2\max(G, v_{\max})$, and both the greedy prefix and the single item of value $v_{\max}$ are feasible. □

Remark 4.3. The fallback is necessary: with $B = 100$ and items $(v_1, c_1) = (2, 1)$, $(v_2, c_2) = (100, 100)$, the greedy prefix is item 1 alone with $G = 2$, and the optimum is item 2 with value 100.

Proposition 4.4 (Equal costs)**.** *If* $c_i = c$ *for all* i*, then reviewing the* $k = \lfloor B/c \rfloor$ *items with the largest* $a_i r_i$ *is exactly optimal for* (2)*.*

Proof. A set is feasible if and only if it has at most k elements. Since every $v_i \ge 0$, the sum of v_i over any set of at most k items is at most the sum of the k largest values. □

5 Certifying the Unreviewed Remainder

The allocation rules rest on (A1). If the machine's confidence is inflated, as Guo et al. [2017] find is typical for uncorrected neural scores, the released set U contains more error than $\sum_{i \in U} a_i r_i$ predicts, and nothing in Sections 3 and 4 reveals it. The remedy does not involve the machine: draw a random sample from U, have it reviewed, and bound the error rate by the exact binomial distribution.

Throughout this section U is a fixed finite set of M released items, of which an unknown number D are wrong, and $p = D/M$ is the *error rate* of U. Nothing is assumed about q. Fix a tolerance $\varepsilon \in (0,1)$ and a risk level $\delta \in (0,1)$.

Theorem 5.1 (Zero-finding certificate)**.** *Draw* n *items independently and uniformly at random from* U *(with replacement) and determine correctly whether each is wrong. Let* X *be the number of wrong items drawn. Then* $\mathbb{P}(X = 0) = (1-p)^n$*. Consequently, if*

$$n \;\ge\; \frac{\ln(1/\delta)}{\ln\!\big(1/(1-\varepsilon)\big)}, \tag{3}$$

then for every population with $p > \varepsilon$ *the probability of finding zero errors is less than* δ*. The simpler condition* $n \ge \ln(1/\delta)/\varepsilon$ *implies* (3)*.*

Proof. Each draw is wrong with probability p independently, so $X \sim \mathrm{Bin}(n, p)$ and $\mathbb{P}(X = 0) = (1-p)^n$. If $p > \varepsilon$ then $(1-p)^n < (1-\varepsilon)^n$, and $(1-\varepsilon)^n \le \delta$ is equivalent to $n \ln(1/(1-\varepsilon)) \ge \ln(1/\delta)$, which is (3). For the last claim, $\ln(1/(1-\varepsilon)) = -\ln(1-\varepsilon) \ge \varepsilon$ for $\varepsilon \in (0,1)$, so $\ln(1/\delta)/\varepsilon \ge \ln(1/\delta)/\ln(1/(1-\varepsilon))$. □

Remark 5.2 (What "confidence $1-\delta$" means)**.** The statement is about the procedure: a rule that certifies when the sample is clean certifies a population with $p > \varepsilon$ with probability below δ. It is not a posterior probability that $p \le \varepsilon$, which would require a prior on p.

When errors are found, the same logic gives the one-sided interval of Clopper and Pearson [1934]. For $0 \le k \le n$ let $F(k; p) = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$ be the binomial distribution function.

Definition 5.3 (Clopper-Pearson upper limit)**.** For $k < n$, $p_{\mathrm{U}}(k)$ is the unique $p \in (0,1)$ with $F(k; p) = \delta$. For $k = n$, $p_{\mathrm{U}}(n) = 1$.

Theorem 5.4 (Exact upper confidence limit)**.** *Fix $n \ge 1$ and $\delta \in (0,1)$.*

(a) *$p_{\mathrm{U}}(k)$ is well defined, and for $k < n$ it equals the $(1-\delta)$ quantile of the Beta distribution with parameters $(k+1,\, n-k)$.*

(b) *If $X \sim \mathrm{Bin}(n, p)$, then $\mathbb{P}(p \le p_{\mathrm{U}}(X)) \ge 1-\delta$ for every $p \in [0,1]$.*

(c) *$p_{\mathrm{U}}(0) = 1 - \delta^{1/n} \le \ln(1/\delta)/n$. For $\delta = 0.05$ this gives $p_{\mathrm{U}}(0) < 3/n$, the "rule of three".*

Proof. (a) Fix $k < n$. Differentiating term by term and using $j\binom{n}{j} = n\binom{n-1}{j-1}$ and $(n-j)\binom{n}{j} = n\binom{n-1}{j}$,

$$\frac{\partial}{\partial p} F(k; p) = n \sum_{j=0}^{k-1} T_j \;-\; n \sum_{j=0}^{k} T_j = -\, n\, T_k, \qquad T_j = \binom{n-1}{j} p^j (1-p)^{n-1-j},$$

which is negative on $(0,1)$. So $F(k; \cdot)$ is continuous and strictly decreasing from $F(k; 0) = 1$ to $F(k; 1) = 0$, and the equation $F(k; p) = \delta$ has exactly one root. Moreover $n\binom{n-1}{k} = n!/(k!\,(n-k-1)!)$ is the normalising constant of the $\mathrm{Beta}(k+1, n-k)$ density $p^k(1-p)^{n-k-1}$, so $1 - F(k; p)$ has the same derivative as the $\mathrm{Beta}(k+1, n-k)$ distribution function and the same value 0 at $p = 0$; the two coincide. Hence $F(k; p) = \delta$ if and only if p is the $(1-\delta)$ quantile.

(b) Fix p. If $X = n$ then $p_{\mathrm{U}}(X) = 1 \ge p$. For $k < n$, strict monotonicity gives $p > p_{\mathrm{U}}(k)$ if and only if $F(k; p) < \delta$. Let $K = \{k < n : F(k; p) < \delta\}$. Since $F(k; p)$ is non-decreasing in k, the set K is either empty or of the form $\{0, \dots, k^*\}$. In the first case $\mathbb{P}(p > p_{\mathrm{U}}(X)) = 0$. In the second, $\mathbb{P}(p > p_{\mathrm{U}}(X)) = \mathbb{P}(X \le k^*) = F(k^*; p) < \delta$.

(c) $F(0; p) = (1-p)^n = \delta$ gives $p = 1 - \delta^{1/n}$. With $x = \ln(1/\delta)/n$, $1 - \delta^{1/n} = 1 - e^{-x} \le x$. For $\delta = 0.05$, $\ln 20 = 2.9957\ldots < 3$. □

For illustration, with $n = 300$ and $\delta = 0.05$ the upper limits are $p_{\mathrm{U}}(0) = 0.0099$, $p_{\mathrm{U}}(1) = 0.0157$, $p_{\mathrm{U}}(2) = 0.0208$ and $p_{\mathrm{U}}(3) = 0.0256$ to four decimal places. Figure 2 plots the required sample size (3).

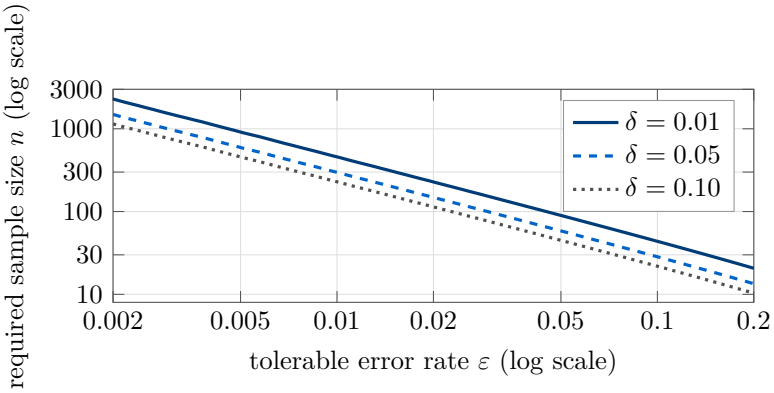


Figure 2: Sample size $n = \ln(1/\delta)/\ln(1/(1-\varepsilon))$ required by Theorem 5.1, against the tolerable error rate ε, for three risk levels δ. Halving the tolerance doubles the sample; tightening the risk level from 0.10 to 0.01 doubles it once. The population size does not enter.

Proposition 5.5 (Finite population)**.** *If the $n \le M$ items are instead drawn uniformly without replacement, then $\mathbb{P}(X = 0) \le (1-p)^n$. The zero-finding certificate of Theorem 5.1 therefore remains valid, with the same n, under sampling without replacement.*

Proof. Without replacement, $\mathbb{P}(X = 0) = \binom{M-D}{n}/\binom{M}{n} = \prod_{j=0}^{n-1} \frac{M-D-j}{M-j}$ when $n \le M - D$, and $\mathbb{P}(X = 0) = 0$ otherwise. For $0 \le j < M$, $(M - D - j)\,M \le (M - D)(M - j)$ because the difference of the two sides is $jD \ge 0$. So each factor is at most $(M - D)/M = 1 - p$, and the product is at most $(1-p)^n$. □

This covers the zero-finding certificate only. When $k > 0$ errors are found, the exact finite-population limit inverts the hypergeometric distribution; I do not claim that $p_{\mathrm{U}}(k)$ is conservative for every k and every population.

Relation to audit sampling and to other guarantees. International Standard on Auditing 530 [International Auditing and Assurance Standards Board, 2009] requires the auditor to determine a sample size sufficient to reduce sampling risk to an acceptably low level. In its vocabulary ε is a tolerable rate of deviation and δ a sampling risk; Theorem 5.1 is one statistical way to meet the requirement, not one the standard prescribes. Wald [1945] showed that a sequential test, which stops as soon as the evidence is decisive, controls both error probabilities with a smaller expected sample than a fixed-size test; the fixed-n certificate is not valid under ad hoc early stopping. Conformal prediction [Vovk et al., 2005, Angelopoulos and Bates, 2021] gives prediction sets that contain the truth with a prescribed marginal probability under exchangeability, and selective classification [Geifman and El-Yaniv, 2017] uses a labelled held-out sample to choose a confidence threshold below which the classifier abstains, with a high-probability bound on the error rate of accepted predictions. All of these need a clean labelled sample exchangeable with the items of interest.

6 Dollar-Weighted Certification

Theorem 5.1 bounds the fraction of *items* in error. In diligence the relevant quantity is the fraction of *dollars*. The two can differ arbitrarily: if U consists of 99 items of USD 1,000 and one item of USD 901,000 and only the large item is wrong, the item error rate is 1% and the dollar error fraction is 90.1%. A uniform sample of 29 items misses the wrong item with probability $0.99^{29} \approx 0.75$.

Definition 6.1 (Dollar error fraction)**.** Let $A = \sum_{i \in U} a_i$. Let $t_i \in [0,1]$ be the *taint* of item i, the fraction of its exposure that is in error; an item that is wholly right has $t_i = 0$ and one that is wholly wrong has $t_i = 1$. The dollar error fraction of U is $\theta = \sum_{i \in U} a_i t_i \,/\, A$.

Theorem 6.2 (Sampling proportional to exposure)**.** *Draw $I_1, \ldots, I_n$ independently from U with $\mathbb{P}(I_s = i) = a_i/A$, and observe the taints t_{I_s} correctly.*

(a) *$\mathbb{E}[t_{I_s}] = \theta$, so the sample mean $\hat{\theta} = \frac{1}{n}\sum_{s=1}^{n} t_{I_s}$ is unbiased for θ.*

(b) *If every item is wholly right or wholly wrong ($t_i \in \{0,1\}$), then the number of wrong items drawn is $\mathrm{Bin}(n, \theta)$, and Theorems 5.1 and 5.4 hold verbatim with θ in place of p.*

Proof. (a) $\mathbb{E}[t_{I_s}] = \sum_{i \in U} (a_i/A)\, t_i = \theta$, and the mean of n such variables has the same expectation.

(b) With $t_i \in \{0,1\}$ each draw is wrong with probability $\sum_{i : t_i = 1} a_i/A = \theta$, independently. □

The all-or-nothing assumption matters only in (b). ISA 530 lists value-weighted selection among its sampling approaches [International Auditing and Assurance Standards Board, 2009]. In the example above, a single exposure-weighted draw hits the wrong item with probability 0.901.

7 The Anchoring Floor

Assumption (A2) says that review resolves the item. A reviewer in a machine-assisted workflow, however, is usually shown the machine's answer and asked whether it is right. Anchoring [Tversky and Kahneman, 1974] and automation bias [Parasuraman and Manzey, 2010] both predict that a wrong answer presented this way is sometimes accepted. The model below takes that as given and asks what follows.

Definition 7.1 (Anchored review)**.** A reviewer of item i sees the machine's decision and produces a final decision. Let $r = \mathbb{P}(W)$ be the machine's error probability on the item. For $r > 0$ let

$$\beta = \mathbb{P}(\text{final decision equals the machine's decision} \mid W)$$

be the *anchor acceptance probability*, and for $r < 1$ let $\gamma = \mathbb{P}(\text{final decision is wrong} \mid W^c)$ be the probability that the reviewer spoils a correct machine decision (read $\beta r = 0$ when $r = 0$ and $\gamma(1-r) = 0$ when $r = 1$).

Theorem 7.2 (Anchoring floor)**.** *Under anchored review the probability R that the final decision is wrong satisfies*

$$R \;\geq\; \beta r + \gamma(1-r) \;\geq\; \beta\, r.$$

The first inequality is an equality if and only if a reviewer who departs from a wrong machine decision always reaches the right one, and the second if and only if $\gamma(1-r) = 0$. In particular, if the reviewer never introduces a new error, then $R = \beta r$ exactly.

Proof. By the law of total probability, $R = \mathbb{P}(\text{final wrong} \mid W)\, r + \gamma(1-r)$. On W the machine's decision is wrong, so the event that the final decision equals the machine's decision is contained in the event that the final decision is wrong; hence $\mathbb{P}(\text{final wrong} \mid W) \geq \beta$, with equality exactly when the reviewer, on departing from a wrong machine decision, is correct with conditional probability one. The second inequality holds because $\gamma(1-r) \geq 0$. □

The floor is linear in the machine's error and independent of how much review is done. Over a reviewed set S with $\beta_i \geq \beta$, the expected residual dollar error is at least $\beta \sum_{i \in S} a_i r_i$.

Remark 7.3 (The reject rule under anchored review)**.** If a reviewed item remains wrong with probability $\beta_i r_i$, it contributes $c_i + \beta_i a_i r_i$ to expected cost when reviewed and $a_i r_i$ when released, so the pointwise comparison of Theorem 3.1 gives the rule $a_i r_i > c_i/(1-\beta_i)$: anchoring raises the price of a review by $1/(1-\beta_i)$.

The natural response to a floor is a second reviewer. Suppose two reviewers each see the machine's decision, and the item is corrected if either rejects it. Condition throughout on W, and let A_1, A_2 be the events that reviewers 1 and 2 accept the wrong anchor, with $\mathbb{P}(A_j \mid W) = \beta_j$. The error survives both reviewers at least on $A_1 \cap A_2$, so the residual error is at least $\mathbb{P}(A_1 \cap A_2 \mid W)\, r$.

Theorem 7.4 (Two anchored reviewers)**.** *Write $\beta_{12} = \mathbb{P}(A_1 \cap A_2 \mid W)$. If A_1 and A_2 are independent given W, then $\beta_{12} = \beta_1 \beta_2$. More generally, suppose there is a random variable Z (the common anchor), describing the item and the wrong answer presented, such that A_1 and A_2 are independent given (W, Z), with $b_j(Z) = \mathbb{P}(A_j \mid W, Z)$. Then*

$$\beta_{12} \;=\; \beta_1 \beta_2 + \operatorname{Cov}\big(b_1(Z),\, b_2(Z) \mid W\big).$$

If Z is real-valued and b_1, b_2 are both non-decreasing, the covariance is non-negative and $\beta_{12} \geq \beta_1 \beta_2$. If the reviewers are exchangeable, $b_1 = b_2 = b$ with $\beta = \mathbb{E}[b(Z) \mid W]$, then

$$\beta^2 \;\leq\; \beta_{12} \;=\; \beta^2 + \operatorname{Var}\big(b(Z) \mid W\big) \;\leq\; \beta,$$

with equality on the left if and only if $b(Z)$ is almost surely constant and on the right if and only if $b(Z) \in \{0, 1\}$ almost surely.

Proof. All probabilities are conditional on W. The first claim is the definition of independence. For the second, by conditional independence,

$$\beta_{12} = \mathbb{E}[b_1(Z) b_2(Z)] = \mathbb{E}[b_1(Z)]\, \mathbb{E}[b_2(Z)] + \operatorname{Cov}(b_1(Z), b_2(Z)),$$

and $\mathbb{E}[b_j(Z)] = \beta_j$. If b_1, b_2 are non-decreasing and Z' is an independent copy of Z, then $(b_1(Z) - b_1(Z'))(b_2(Z) - b_2(Z')) \geq 0$ pointwise, and its expectation is $2 \operatorname{Cov}(b_1(Z), b_2(Z))$,

which is therefore non-negative. For exchangeable reviewers the covariance is $\mathrm{Var}(b(Z)) \geq 0$, zero if and only if $b(Z)$ is almost surely constant. Finally $b(Z)^2 \leq b(Z)$ because $0 \leq b \leq 1$, so $\beta_{12} = \mathbb{E}[b(Z)^2] \leq \mathbb{E}[b(Z)] = \beta$, with equality if and only if $b(Z)^2 = b(Z)$ almost surely. □

A wrong answer that looks plausible looks plausible to both reviewers. As an illustrative calculation, suppose half of the machine's errors are plausible, accepted by each reviewer with probability 0.5, and half are obvious, accepted with probability 0.1. Each reviewer has $\beta = 0.3$ and independence would predict a joint miss rate of 0.09; the common-anchor model gives $\beta_{12} = \frac{1}{2}(0.25) + \frac{1}{2}(0.01) = 0.13$.

Remark 7.5 (Design consequence)**.** In the allocation step anchored review is tolerable at the price given in Remark 7.3. In the certification step it is not: the sample exists to measure the machine's error, and an anchored reviewer undercounts it. The certification sample should be reviewed *blind*: the reviewer derives the answer from source without seeing the machine's output, and the two are compared afterwards.

Theorem 7.6 (Certification with imperfect detection)**.** *In the setting of Theorem 5.1, suppose that each wrong item drawn is flagged with probability at least $1 - \beta_0$, independently across draws, where $\beta_0 \in [0, 1)$. Then the probability that no item is flagged is at most $\big(1 - (1 - \beta_0)\, p\big)^n$, and a sample with no flag certifies $p \leq \varepsilon$ at confidence $1 - \delta$ provided*

$$n \;\geq\; \frac{\ln(1/\delta)}{\ln\!\big(1/(1 - (1 - \beta_0)\varepsilon)\big)}. \tag{4}$$

With the sample size (3) computed as if $\beta_0 = 0$, a sample with no flag certifies only $p \leq \varepsilon/(1 - \beta_0)$.

Proof. Each draw is wrong with probability p and, if wrong, is flagged with probability at least $1 - \beta_0$, so each draw independently produces a flagged error with probability at least $(1 - \beta_0)p$. The probability of no flagged error in n draws is at most $(1 - (1 - \beta_0)p)^n$; flags raised on correct items can only reduce the probability of a sample with no flag. If $p > \varepsilon$ this is less than $(1 - (1 - \beta_0)\varepsilon)^n$, which is at most δ under (4). For the last claim, apply the first with ε replaced by $\varepsilon' = \varepsilon/(1 - \beta_0)$ when $\varepsilon' < 1$: the right-hand side of (4) for ε' equals the right-hand side of (3) for ε. □

The required sample grows by approximately the factor $1/(1 - \beta_0)$. More importantly, β_0 cannot be estimated from an anchored sample itself, because the errors the reviewer missed are exactly the ones not recorded.

8 Reviewer Quality without Ground Truth

Ground truth about reviewers is usually unavailable, but overlapping assignments are cheap. Dawid and Skene [1979] posit that each item has an unobserved true class, that each observer has an individual matrix of error rates, and that observers' responses are conditionally independent given the true class; they estimate the error rates and class prevalences by maximum likelihood using the EM algorithm. The conditional-independence assumption is

the one Theorem 7.4 warns about: plausible for blind reviewers, implausible for reviewers who share an anchor.

A simpler summary is agreement. Cohen [1960] pointed out that raw agreement between two raters includes the agreement expected by chance and proposed the coefficient $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is the observed proportion of agreement and p_e is the agreement expected if the raters labelled independently with their observed marginal frequencies. In review of machine output one class ("correct") dominates, the regime in which raw agreement is least informative.

Proposition 8.1 (Raw agreement overstates reliability under class imbalance)**.** *Two reviewers label each item "correct" or "error", independently of each other, each saying "correct" with probability m. Then $\kappa = 0$, while the probability of agreement is $p_o = m^2 + (1-m)^2 = 1 - 2m(1-m)$, which tends to 1 as $m \to 1$. For $m = 0.95$, $p_o = 0.905$.*

Proof. Under independence the reviewers agree with probability $m^2 + (1-m)^2$, which is also p_e by definition, so $p_o = p_e$ and $\kappa = 0$. At $m = 0.95$, $1 - 2(0.95)(0.05) = 0.905$. □

A finite table makes the same point exactly. Suppose two reviewers examine 100 items: on 90 both say "correct", on 5 only the first says "error", on 5 only the second does, and on none do both. Then $p_o = 0.90$, each reviewer says "correct" on 95 items, $p_e = 0.95^2 + 0.05^2 = 0.905$, and $\kappa = (0.900 - 0.905)/(1 - 0.905) = -1/19 \approx -0.053$. Ninety per cent agreement coexists with agreement slightly worse than chance: the reviewers never once identified the same error.

9 The Release Gate

The pieces compose into a gate.

Definition 9.1 (Release gate)**.** Fix a tolerance ε, a risk level δ and a bound $\beta_0 \in [0, 1)$ on the blind reviewers' miss probability. *Allocate*: choose a review set S by Theorem 3.1, Theorem 4.2 or any other rule, review it, and let U be the remainder. *Sample*: draw n items from U uniformly at random, with n satisfying (4), by a mechanism independent of which items are wrong, and have each reviewed blind. *Decide*: if no sampled item is flagged, release; otherwise hold, correct the flagged items, revise the allocation, and repeat with a fresh sample.

Theorem 9.2 (End-to-end guarantee)**.** *Suppose that in each pass through the gate* (G1) *the sample is drawn uniformly at random from the current remainder U, independently of which items of U are wrong given everything that happened earlier, and* (G2) *each wrong item drawn is flagged with probability at least $1 - \beta_0$, independently across draws. Let pass j use risk level δ_j and sample size (4) with $\delta = \delta_j$, and let p_j be the error rate of the remainder at pass j. Then, for any allocation rule whatever and any values of the confidences q_i, calibrated or not,*

$$\mathbb{P}(\textit{the gate releases at some pass } j \le J \textit{ with } p_j > \varepsilon) \;\le\; \sum_{j=1}^{J} \delta_j.$$

Table 1: Illustrative example (amounts in USD): the dollar-weighted rule of Theorem 3.1 against a global confidence cut-off of 0.90. $v_i = a_i(1-q_i)$; $\tau_i = 1 - c_i/a_i$, to four places. The cost columns give each item's contribution to (1); **R** marks a reviewed item.

						Dollar-weighted		Global 0.90	
i	a_i	q_i	c_i	v_i	τ_i		cost		cost
1	500,000	0.970	600	15,000	0.9988	**R**	600		15,000
2	250,000	0.990	600	2,500	0.9976	**R**	600		2,500
3	120,000	0.950	400	6,000	0.9967	**R**	400		6,000
4	60,000	0.920	400	4,800	0.9933	**R**	400		4,800
5	40,000	0.995	400	200	0.9900		200		200
6	15,000	0.850	300	2,250	0.9800	**R**	300	**R**	300
7	8,000	0.980	300	160	0.9625		160		160
8	4,000	0.880	300	480	0.9250	**R**	300	**R**	300
9	1,200	0.800	300	240	0.7500		240	**R**	300
10	900	0.700	300	270	0.6667		270	**R**	300
Expected total cost							3,470		29,860

In particular J passes at level δ/J each keep the overall risk at δ. If the sample is instead drawn in proportion to exposure, the same statement holds with the dollar error fraction θ_j in place of p_j, for items that are wholly right or wholly wrong.

Proof. Fix a pass j and condition on the history $\mathcal{H}_j$ of everything before the sample of pass j is drawn: the allocation, all earlier samples and their outcomes. Given $\mathcal{H}_j$ the remainder and its error rate p_j are fixed. On the event $p_j > \varepsilon$, assumptions (G1) and (G2) and Theorem 7.6 give $\mathbb{P}(\text{no flag at pass } j \mid \mathcal{H}_j) \le \delta_j$. Hence

$$\mathbb{P}(\text{release at pass } j,\ p_j > \varepsilon) = \mathbb{E}\big[\mathbf{1}\{p_j > \varepsilon\}\,\mathbb{P}(\text{no flag at pass } j \mid \mathcal{H}_j)\big] \le \delta_j,$$

and the union bound over $j \le J$ gives the claim. No property of the allocation or of q was used. For exposure-weighted sampling, each draw is wrong with probability θ_j by Theorem 6.2(b), and the proof of Theorem 7.6 goes through with θ_j in place of p. □

Remark 9.3 (What the guarantee leaves to the allocation)**.** The gate is safe for any allocation but only *passes* for a good one. A correct blind sample from a remainder with error rate p is clean with probability $(1-p)^n$. With $\varepsilon = 0.01$, $\delta = 0.05$ and $n = 299$, a remainder with $p = 0.001$ passes with probability $0.999^{299} \approx 0.74$ and one with $p = 0.005$ with probability about 0.22. Miscalibration costs rework and delay; under (G1) and (G2) it does not cost a wrong release beyond the stated risk.

10 An Illustrative Worked Example

Table 1 lists ten hypothetical items; the numbers are illustrative and not drawn from any engagement.

Table 2: Smallest integer n with $n \geq \ln(1/\delta)/\ln(1/(1-\varepsilon))$ (Theorem 5.1); in parentheses, the smallest integer with $n \geq \ln(1/\delta)/\varepsilon$.

	$\varepsilon = 0.10$	$\varepsilon = 0.05$	$\varepsilon = 0.02$	$\varepsilon = 0.01$	$\varepsilon = 0.005$
$\delta = 0.10$	22 (24)	45 (47)	114 (116)	230 (231)	460 (461)
$\delta = 0.05$	29 (30)	59 (60)	149 (150)	299 (300)	598 (600)
$\delta = 0.01$	44 (47)	90 (93)	228 (231)	459 (461)	919 (922)

Unconstrained allocation. The dollar-weighted rule reviews items $1, 2, 3, 4, 6, 8$ at a cost of 2,600 and releases items $5, 7, 9, 10$ with expected error $200 + 160 + 240 + 270 = 870$, a total of 3,470. The global cut-off at 0.90 reviews items $6, 8, 9, 10$ for 1,200 and releases the others with expected error $15{,}000 + 2{,}500 + 6{,}000 + 4{,}800 + 200 + 160 = 28{,}660$, a total of 29,860. Reviewing nothing costs 31,900 and reviewing everything 3,900, so the customary cut-off captures 2,040 of an available improvement of 28,430, about 7%. It fails in both directions: it spends 600 reviewing items 9 and 10, whose combined expected error is 510, and it releases item 1 at confidence 0.97 with an expected error of 15,000. Tuning helps only by reviewing nearly everything: the best global rule sets τ just above 0.99, reviews every item except item 5, and costs $3{,}500 + 200 = 3{,}700$, which is $230 = (300 - 240) + (300 - 270) + (300 - 160)$ above the optimum. Any cut-off at or below 0.97 releases item 1 and costs at least 19,860.

Budgeted allocation. Cap review at $B = 1{,}500$. The densities v_i/c_i of items 1 to 10 are 25, 4.17, 15, 12, 0.5, 7.5, 0.53, 1.6, 0.8 and 0.9. The greedy prefix is items $1, 3, 4$ at cost 1,400, removing $G = 25{,}800$ of expected error; item 6 is the break item. Lemma 4.1 gives $F = 25{,}800 + \frac{100}{300} \cdot 2{,}250 = 26{,}550$, so the greedy set is within 750 of optimal without enumeration, and enumeration of the 2^{10} subsets confirms that it is optimal.

Certification. By Table 2, a remainder of any size is certified to an error rate of at most 1% at 95% confidence by 299 clean blind reviews. Had those reviews been anchored with $\beta_0 = 0.3$, Theorem 7.6 says the same clean sample certifies only $p \leq 0.01/0.7 \approx 1.43\%$.

11 Implications for Practice

Three design rules follow. Rank the review queue by expected dollar error per unit cost, $a_i r_i / c_i$, not by confidence. Separate review that corrects, which can tolerate anchoring at a known price, from review that measures, which must be blind. Report the certificate with its denominator: n, the errors found, ε, δ, and whether sampling was by item or by exposure.

These rules are consistent with how Finsider's platform is organised. A company's results are held behind a release gate until verification passes, with human reviewers in the loop; and any figure extracted by an AI layer is kept separate from the deterministic core computed from ledger data, is labelled as needing verification, and does not feed headline metrics until verified. Read through Theorem 3.1, an item that feeds a headline metric has a large exposure, so the optimal rule reviews it at essentially any machine confidence. I make no claim here about the sampling plans or thresholds used in any particular engagement.

12 Falsifiable Predictions

These are predictions about future controlled studies, not reports of any.

Prediction 1 (A positive anchoring floor)**.** In an experiment that seeds known errors into machine output and randomises reviewers between seeing the machine's answer and working blind, seeded errors survive anchored review at a higher rate than they survive blind review. *Refuted if* the two rates are statistically indistinguishable, or anchored reviewers leave fewer errors.

Prediction 2 (Positive dependence between anchored reviewers)**.** When two reviewers separately check the same anchored items, the fraction of seeded errors missed by both exceeds the product of their individual miss rates. *Refuted if* the joint miss rate is at or below that product, within sampling error, in an adequately powered study.

Prediction 3 (Dollar weighting beats a tuned global cut-off)**.** On ledger populations whose exposures span at least two orders of magnitude, at equal review effort, ordering the queue by $a_i r_i / c_i$ leaves fewer realised dollars in error than the best single confidence cut-off chosen in hindsight. *Refuted if* the hindsight-optimal cut-off matches or beats the dollar-weighted ordering in a majority of such populations.

Prediction 4 (Confidence understates error after a shift)**.** When a model calibrated on one set of companies is applied to a company with a materially different chart of accounts, the error count predicted by $\sum_{i \in U} r_i$ falls below a two-sided 95% exact binomial interval from a blind sample of U more often than the nominal 2.5%. *Refuted if* it does so in at most 2.5% of such deployments, within sampling error.

13 Limitations

Confidence. Theorems 3.1 to 4.2 assume (A1). If q_i is miscalibrated the allocation is suboptimal and nothing in those theorems detects it; the gate remains safe but may rarely pass (Remark 9.3). I have not analysed how to recalibrate q from gate outcomes.

Review. The anchoring model is minimal and ignores fatigue, learning and time pressure. Theorem 7.2 remains a lower bound under such refinements, but the value of β is an empirical quantity this paper does not supply.

Exchangeability. Assumption (G1) excludes convenience samples, samples chosen by the reviewer, and samples drawn before the remainder is final. The certificate bounds error in what was processed, not completeness of what was provided. Assumption (G2) fails if a reviewer misses every instance of one error type; β_0 must then be a worst case over types.

Additivity. Exposure is treated as known and losses as additive. Real errors can offset or compound, reviewing one item can be informative about others, and partial errors are handled only through the unbiasedness in Theorem 6.2(a).

14 Conclusion

A human should see the items whose expected dollar error exceeds the cost of looking, which makes the confidence threshold a function of exposure. The items no human sees can be

certified by a blind random sample whose size depends only on the tolerance and the risk level. Review that begins from the machine's answer leaves behind a fixed share of the machine's errors. Composed, these give a release gate whose guarantee can be written down with its assumptions (uniform sampling, blind and accurate review of the sample, meaningful confidence for allocation), and those are where a real process can fail.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

George B. Dantzig. Discrete-variable extremum problems. *Operations Research*, 5(2):266–288, 1957.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

International Auditing and Assurance Standards Board. International standard on auditing 530: Audit sampling. IFAC, 2009.

Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.