Validation Transfer and Reverse Winner's Curse in AI Advice

Christian Roessler*

May 31, 2026

Abstract

When AI users observe performance on familiar tasks and infer reliability on unfamiliar tasks, "validation transfer" occurs. Associated systematic errors are related to winner's curse logic. This paper studies validation transfer in a linear-quadratic advice model. A user's own understanding of a task is a signal about population-level knowledge, and hence the training support of the AI. The results imply a shading logic: users should reduce AI authority, and may need more human verification, when lack of familiarity indicates weak information support. The main theorem decomposes the welfare loss from transferring authority across task distributions into hidden-heterogeneity and distribution-transfer losses. Conditional validation has Blackwell value; pooled learning can still hide reliability heterogeneity.

JEL: D83, D81, C44, C53, O33.

Keywords: AI advice; validation transfer; reverse winner's curse; information partitions; misspecified learning.

*Department of Economics, California State University, East Bay. Email: christian.roessler@csueastbay.edu. No financial support or potential conflicts to acknowledge.

1 Introduction

Clever Hans was a horse in early 20th century Berlin who solved arithmetic problems by tapping his hoof when the count reached the right answer. Investigations did not uncover any fraud. Instead, the horse responded to cues the audience provided inadvertently. He sensed the onlookers' anticipation at the moment the right number came up, and he had learned that there was an associated reward (Pfungst, 1911). His apparent "knowledge" depended entirely on the information present around him. When the audience knew the answer, the cues were useful. If the problem were genuinely difficult, then Clever Hans could not have solved it either. Yet, the audience might not have realized it, if it believed the horse had miraculous powers.

AI advice has similar features. A user who is impressed with high performance on tasks that are common and benchmarked may rely on AI to execute a task that is unfamiliar and rare. One issue is that observed reliability does not necessarily carry over to the new problem. Additionally, personal familiarity with the problem is an informative signal, since it is correlated with the data environment behind the AI output, namely what human knowledge could have trained or validated it.

How should validation evidence be used when AI reliability varies across task types? The paper proceeds from the central idea that personal familiarity contains common-value information. A task familiar to one person is more likely to be familiar to other people whose data or work feeds into the AI system. A validation sample drawn from one task distribution can therefore induce an authority weight that is wrong for another task distribution. Conditional validation may have value even when the pooled validation estimate is learned without bias.

There is a *reverse winner's curse* at play. In a standard common-value auction, a winning bidder should infer that his signal is high relative to the signals of other bidders (Capen, Clapp, and Campbell, 1971; Milgrom and Weber, 1982; Thaler, 1988). Here a user who finds a task familiar should infer that other people may also find it familiar. Conversely, a

user who finds a task unfamiliar should consider that the same may apply to other people, limiting the information that is available to the AI system. Overconfidence in this context is treating familiar-domain validation as if it were neutral evidence about unfamiliar tasks.

The paper connects information economics with the economics of AI-assisted decisions. The central object is not a prediction, treatment effect, or belief, but the authority weight assigned to AI advice. A user observes validation evidence, learns about the average performance of the system, and then transfers that information to a deployment environment that may contain a different mix of tasks. The key question is whether a validation-based authority rule remains appropriate when AI reliability varies systematically across task types.

This perspective relates to several strands of research. In the economics of AI, Agrawal, Gans, and Goldfarb (2018) and Agrawal, Gans, and Goldfarb (2019) emphasize the distinction between prediction and judgment, while Kleinberg et al. (2015), Mullainathan and Spiess (2017), Athey and Imbens (2019), and Athey and Wager (2021) study machine predictions as inputs into economic decisions. The present paper shifts attention from the quality of predictions to the choice of how much authority those predictions should receive. The optimal authority weight depends on the task type, implying that validation evidence from one environment may not transfer cleanly to another.

The paper is also related to the literature on reliance on algorithmic advice. Early work distinguished use, misuse, disuse, and trust in automation (Parasuraman and Riley, 1997; Lee and See, 2004). More recent studies document algorithm aversion, algorithm appreciation, willingness to rely on imperfect algorithms, and human-AI complementarity (Dietvorst, Simmons, and Massey, 2015, 2018; Logg, Minson, and Moore, 2019; Steyvers et al., 2022; Vaccaro, Almaatouq, and Malone, 2024). Constant authority rules are additionally related to forecast-combination weights (Bates and Granger, 1969; Clemen, 1989). The closest paper is Balakrishnan, Ferreira, and Tong (2026), which studies situations in which private information calls for instance-specific adjustment of a constant advice weight. Here the authority weight itself is learned from validation evidence, and the main issue is whether that learned

weight should be transferred across task distributions. Relatedly, Agarwal, Moehring, and Wolitzky (2025) study how information about both humans and AI systems should shape decision design.

The transfer problem places the paper in conversation with the literature on external validity and transportability. Hotz, Imbens, and Mortimer (2005), Manski (2004), Andrews and Oster (2019), Meager (2019), and Vivalt (2020) study when evidence obtained in one environment can be used in another, while Pearl and Bareinboim (2014) formalize transportability using causal structure. Andrews et al. (2025) examine the transfer of economic models across domains. The present paper studies the transfer of a decision rule rather than a treatment effect or forecasting model. Even when validation delivers a correct average estimate, that estimate may be the wrong object for a heterogeneous deployment problem.

The analysis is particularly relevant for AI systems because reliability often varies across domains. Recent works examine how people generalize language-model performance across tasks (Vafa, Rambachan, and Mullainathan, 2024), how validation and leakage affect language-model measurement (Ludwig, Mullainathan, and Rambachan, 2026), and how AI performance can be highly uneven across realistic tasks (Dell'Acqua et al., 2026). A large machine-learning literature looks at domain adaptation, dataset shift, shortcut learning, benchmark validity, and model documentation (Quiñonero-Candela et al., 2009; Ben-David et al., 2010; Recht et al., 2019; Koh et al., 2021; Geirhos et al., 2020; Lapuschkin et al., 2019; Liang et al., 2023; Mitchell et al., 2019; Gebru et al., 2021; Raji et al., 2021). These contributions explain why pooled benchmark performance may be fragile. This paper adds a decision-theoretic layer by characterizing the welfare consequences of transferring authority weights across task distributions.

Methodologically, the paper combines ideas from information economics and misspecified learning. The value of conditional validation is characterized through a Blackwell comparison (Blackwell, 1953), linking the analysis to the broader information-design literature (Kamenica and Gentzkow, 2011). The learning results draw on the literature on convergence

under misspecification (Berk, 1966; White, 1982; Esponda and Pouzo, 2016, 2021). A user can learn the pseudo-true pooled projection and still use the wrong object on particular task types.

The model is a linear-quadratic advice problem. A task type determines how informative the AI signal is. A validation distribution generates the authority weight that the user assigns to the AI signal. A deployment distribution determines the tasks on which the weight is used. The main theorem decomposes the welfare loss from validation transfer into two parts. The first part is the loss from using one authority weight when different deployment tasks call for different weights. The second part is the loss from using a weight learned under the validation distribution rather than the best common weight for the deployment distribution.

The supporting results give content to the decomposition. A reverse winner's curse theorem microfounds why familiar tasks should have better expected AI support. A data-processing lemma formalizes the Clever Hans idea: under a specified information environment, the AI output is bounded by the information in that environment. A Blackwell comparison shows that conditional validation and task-familiarity signals have nonnegative ex ante value when those signals refine the information partition. A deployment result separates using more AI authority than the first best from being worse off than without AI. A selective-use implication shows why high-uncertainty tasks may attract AI consultation.

Section 2 develops the reverse winner's curse microfoundation and the Clever Hans bound. Section 3 develops a two-type benchmark. Section 4 presents the general second-moment model. Section 5 proves the validation-transfer theorem. Section 6 studies conditional validation. Section 7 studies learning from pooled validation. Section 8 compares AI use with no AI and discusses selective use. Section 9 concludes. The appendix collects supplementary results on detecting omitted heterogeneity, the limits of typewise monotonicity, and the familiar-side share of aggregate loss.

2 Reverse winner's curse and Clever Hans validation

Before introducing the decision model, it is useful to formalize two ideas that motivate the analysis. The first explains why a person's own familiarity with a task contains information about the reliability of AI advice on that task. The second formalizes the intuition behind the Clever Hans analogy from the introduction.

The common theme is that AI performance depends on the information environment from which the system draws. Tasks that are familiar to many people are more likely to be represented in training data, validation data, and related human knowledge. Tasks that are unfamiliar to a user are more likely to be unfamiliar to the broader population as well. As a result, personal familiarity is informative about the support available to the AI system.

This section establishes these ideas in two steps. First, Theorem 1 shows that familiarity shifts beliefs about the amount of population knowledge available to the AI. This result provides the microfoundation for the reverse winner's curse. Second, Lemma 1 shows that AI outputs cannot contain more information than the underlying information environment from which they are generated. Together, these results explain why validation evidence from familiar tasks may not transfer cleanly to unfamiliar tasks.

2.1 Correlated familiarity

Fix $N \geq 1$. Let $\lambda \in (0,1)$ be drawn from a prior distribution and denote latent task commonness. Conditional on λ, each potential data contributor or audience member $j = 0, 1, \ldots, N$ is familiar with the task with probability λ, independently across j. The decision maker is indexed by 0. Let

$$F_j \in \{0,1\}$$

indicate whether person j is familiar with the task, and let

$$K = \sum_{j=1}^{N} F_j$$

be the amount of population familiarity available to the AI's training environment. AI error variance is a decreasing function of training support:

$$v = v(K), \qquad v(k+1) \leq v(k).$$

The monotonicity condition means that more support in the data environment weakly lowers expected error. The binomial structure is a sufficient condition for the result. The decision model below only uses the implication that familiarity shifts expected AI reliability.

The first result turns this intuition into a stochastic-ordering statement. The decision maker's own familiarity is evidence about the latent commonness of the task, and therefore about how much population knowledge may be available to the AI system.

Theorem 1 (Reverse winner's curse)**.** *In the correlated-familiarity model, the conditional distribution of K given $F_0 = 1$ first-order stochastically dominates the conditional distribution of K given $F_0 = 0$. Consequently, for any decreasing error function $v(K)$,*

$$\mathbb{E}[v(K) \mid F_0 = 1] \leq \mathbb{E}[v(K)] \leq \mathbb{E}[v(K) \mid F_0 = 0].$$

If the posterior distributions of K differ and $v(K)$ is not almost surely constant on their combined support, the inequalities are strict.

Proof. Conditional on λ, F_0 is Bernoulli with success probability λ. Bayes' rule implies that the posterior measure after $F_0 = 1$ has Radon–Nikodym derivative proportional to λ with respect to the prior measure. The posterior measure after $F_0 = 0$ has Radon–Nikodym derivative proportional to $1 - \lambda$. The likelihood ratio between the two posterior measures is proportional to

$$\frac{\lambda}{1-\lambda},$$

which is increasing in λ on $(0, 1)$. Hence $\lambda \mid F_0 = 1$ first-order stochastically dominates $\lambda \mid F_0 = 0$.

Conditional on λ, the training-support variable $K = \sum_{j=1}^{N} F_j$ has binomial distribution with parameters N and λ. This distribution is stochastically increasing in λ. Therefore $K \mid F_0 = 1$ first-order stochastically dominates $K \mid F_0 = 0$.

If $v(k)$ is decreasing, stochastic dominance gives

$$\mathbb{E}[v(K) \mid F_0 = 1] \leq \mathbb{E}[v(K) \mid F_0 = 0].$$

The unconditional expectation is a convex combination of the two conditional expectations, which gives

$$\mathbb{E}[v(K) \mid F_0 = 1] \leq \mathbb{E}[v(K)] \leq \mathbb{E}[v(K) \mid F_0 = 0].$$

The stated strictness condition gives strict inequality in the stochastic-dominance comparison and hence in the expectations. □

The theorem supplies the paper's first formal reason for treating familiarity as evidence. A user who finds a task familiar should infer that the broader environment is more likely to contain related knowledge, examples, and validation opportunities. The proof is a Bayesian updating argument. Observing familiarity tilts the posterior distribution of latent commonness upward; higher commonness makes the distribution of population support larger in the sense of first-order stochastic dominance; a decreasing error function then converts higher support into lower expected AI error.

This is the reverse of the winner's curse logic. In the usual winner's curse, a bidder who wins should worry that winning is bad news: it may mean that his estimate was unusually optimistic relative to others' information. Here, familiarity is also a signal about others' information, but it points in the opposite direction. When a user finds a task familiar, it is part of a universe in which many people may have seen similar problems, produced relevant examples, corrected mistakes, and generated data from which the AI system can benefit. Unfamiliarity has the opposite implication. A task that feels unfamiliar to the user may also be thinly represented in the broader information environment. The theorem therefore

turns a user's own familiarity into a common-value signal about AI reliability. The mistake it warns against is treating observed AI success on familiar tasks as neutral evidence about performance on tasks whose informational support may be much weaker.

2.2 The Clever Hans bound

Let θ denote the payoff-relevant state that the user would like to learn. Let Z denote the information in the relevant audience, training corpus, historical precedent set, or social data environment. Let S denote the AI output. Let X denote the information already held by the user. The Clever Hans condition is

$$S \perp\!\!\!\perp \theta \mid (Z, X).$$

Given the information environment and the user's information, the AI output contains no additional independent information about the state. If the system uses an external tool, database, sensor, or retrieval source, that source belongs in Z.

The next result makes the Clever Hans analogy precise. The AI system's informational content is limited by the environment from which it draws. If the environment contains little independent information about the state, the AI output cannot create that information by itself. For random variables U and V, $I(U;V)$ denotes their mutual information, the standard information-theoretic measure of how much observing one variable reduces uncertainty about the other.[1]

Lemma 1 (Clever Hans information bound)**.** *If* $S \perp\!\!\!\perp \theta \mid (Z, X)$*, then*

$$I(\theta; S \mid X) \leq I(\theta; Z \mid X).$$

Proof. The conditional independence assumption says that, given the user's information X,

[1]For discrete random variables, $I(U;V) = \sum_{u,v} p(u,v) \log\left(p(u,v)/(p(u)p(v))\right)$. Equivalently, $I(U;V) = H(U) - H(U \mid V)$.

the variables form the conditional Markov chain

$$\theta \longrightarrow Z \longrightarrow S.$$

The conditional data-processing inequality states that if, conditional on a variable C, the variables form the Markov chain $A \to B \to D$, then processing B into D cannot increase information about A:

$$I(A; D \mid C) \leq I(A; B \mid C).$$

Applying the inequality with $A = \theta$, $B = Z$, $D = S$, and $C = X$ gives

$$I(\theta; S \mid X) \leq I(\theta; Z \mid X).$$

□

The lemma says, effectively, that the AI output is a garbling of the information environment, conditional on the user's information. The system may aggregate or transform information in Z. It cannot contain more information about θ than the environment from which it is generated.

3 A two-type benchmark

This section gives a small model before the general analysis. The purpose is to isolate the main forces with little notation. There are two task types, familiar tasks F and unfamiliar tasks U. The payoff state is $\theta = \mu_0 + x$, where x is the residual payoff-relevant state after the user's ordinary information. The AI signal is centered and conditionally unbiased for that residual state:

$$y = x + \varepsilon_T, \qquad T \in \{F, U\}.$$

The signal error is unbiased and orthogonal to this residual state. For each $\tau \in \{F, U\}$,

$$\mathbb{E}[x \mid T = \tau] = 0, \qquad \mathrm{Var}(x \mid T = \tau) = A > 0, \qquad \mathbb{E}[\varepsilon_T \mid T = \tau] = 0, \qquad \mathrm{Cov}(x, \varepsilon_T \mid T = \tau) = 0.$$

The AI error variance is type dependent:

$$\mathrm{Var}(\varepsilon_T \mid T = F) = v_F, \qquad \mathrm{Var}(\varepsilon_T \mid T = U) = v_U, \qquad 0 < v_F < v_U.$$

Let q be the unfamiliar-task share in the validation distribution Q, and let p be the unfamiliar-task share in the deployment distribution P. The average error variances under the two distributions are

$$\bar{v}_Q = (1 - q)v_F + qv_U, \qquad \bar{v}_P = (1 - p)v_F + pv_U.$$

As in the general model below, the decision maker has quadratic payoff $u(a, \theta) = a\theta - ca^2/2$ with $c > 0$. Given an authority weight w, the decision maker forms the forecast $\mu_0 + wy$ and chooses $a_w = (\mu_0 + wy)/c$. For each task type τ, the user chooses w to maximize

$$\mathbb{E}\left[a_w\theta - \frac{c}{2}a_w^2 \mid T = \tau\right], \qquad a_w = \frac{\mu_0 + wy}{c}.$$

Solving this problem gives

$$w_\tau = \frac{\mathrm{Cov}(x, y \mid T = \tau)}{\mathrm{Var}(y \mid T = \tau)} = \frac{A}{A + v_\tau}.$$

Thus the optimal authority weights on the two task types are

$$w_F = \frac{A}{A + v_F}, \qquad w_U = \frac{A}{A + v_U}.$$

Equivalently, w_τ is the linear projection coefficient of the residual payoff state x on the AI

signal y, conditional on task type τ. AI authority is higher when the signal contains more payoff-relevant variation and lower when the signal contains more noise. The user should then put less authority on noisier AI advice. Since $v_U > v_F$, the unfamiliar-task weight is lower than the familiar-task weight.

A pooled validation report from Q induces

$$w_Q = \frac{A}{A + \bar{v}_Q}.$$

If $0 < q < 1$, then $w_F > w_Q > w_U$. Pooled validation creates two mistakes at the same time. It underweights AI advice on familiar tasks and overweights AI advice on unfamiliar tasks. The same scalar validation report can create both disuse and misuse. Appendix C records a binary decomposition showing that either side can account for most of the aggregate first-best deficit.

When validation and deployment differ, the best common deployment weight is

$$w_P = \frac{A}{A + \bar{v}_P}.$$

The loss from using w_Q under deployment distribution P has two parts:

$$W_P^{FB} - W_P(w_Q) = \frac{A^2}{2c}\left\{\mathbb{E}_P\left[\frac{1}{A + v(T)}\right] - \frac{1}{A + \bar{v}_P}\right\} + \frac{A + \bar{v}_P}{2c}(w_Q - w_P)^2.$$

The first term is the loss from hiding heterogeneity inside the deployment distribution. The second term is the loss from applying the wrong average, because validation was drawn from Q while deployment is drawn from P.

The numerical example in Table 1 uses $A = c = 1$, $v_F = .25$, $v_U = 4$, $q = .20$, and $p = .80$. Validation is mostly familiar, while deployment is mostly unfamiliar. Let $G_\tau(w) := W_\tau(w) - W_\tau(0)$ denote the welfare gain from using authority weight w rather than no AI on type τ. The example illustrates three formal results below: bidirectional misreliance, the

validation-transfer decomposition, and the distinction between overuse and harm relative to no AI.

Quantity	Value	Interpretation
w_F	.800	optimal authority on familiar tasks
w_U	.200	optimal authority on unfamiliar tasks
w_Q	.500	authority learned from validation
w_P	.235	best common authority for deployment
$G_F(w_Q)$	.344	gain from AI on familiar tasks
$G_U(w_Q)$	$-.125$	gain from AI on unfamiliar tasks
$W_P(w_Q) - W_P(0)$	$-.031$	aggregate gain from transferred authority
Within-deployment loss	.042	loss from hidden heterogeneity under P
Transfer loss	.149	loss from using Q rather than P

Table 1: A two-type numerical example.

The table illustrates that the pooled validation weight lies between the two type-specific weights. Also, the transferred weight is high enough to make AI use harmful on unfamiliar tasks, even though it remains beneficial on familiar tasks. Finally, most of the total first-best loss in this example comes from transfer across distributions rather than from within-deployment pooling.

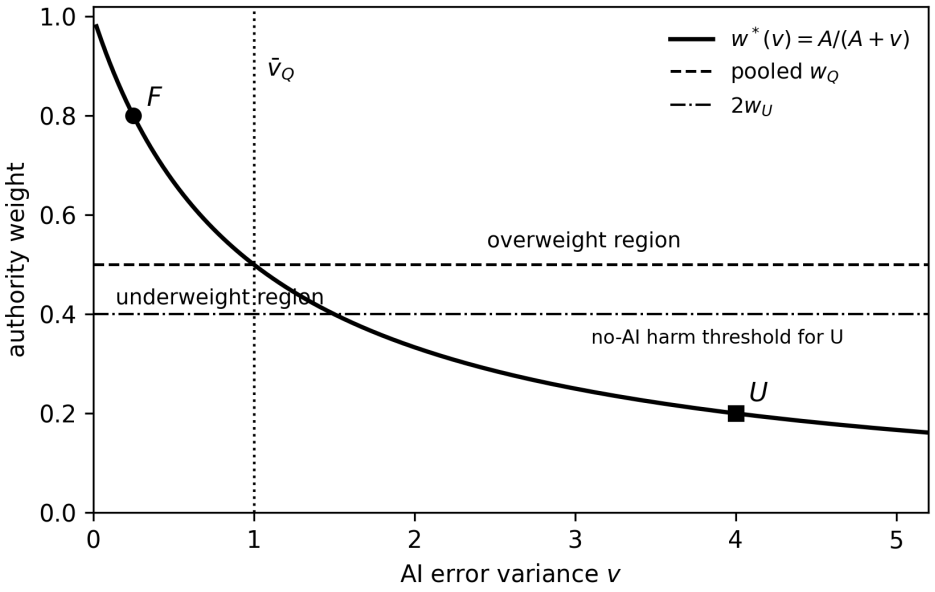


Figure 1: Authority weights in the two-type benchmark. The validation weight lies between the familiar and unfamiliar optima.

Figure 1 makes the same comparison graphically. The authority weight falls as error variance rises. A pooled validation weight must lie between the type-specific weights when validation contains both types.

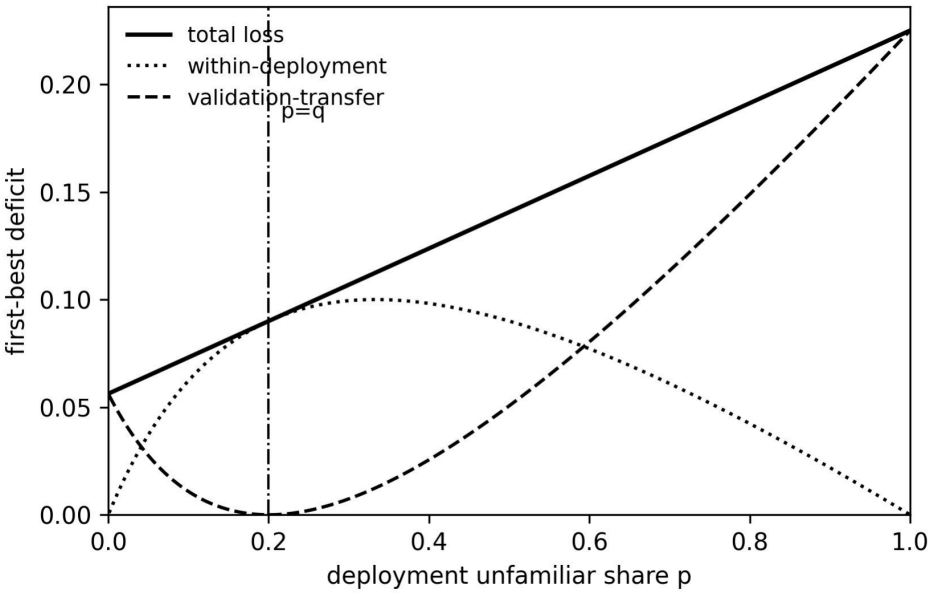


Figure 2: Loss decomposition in the two-type benchmark. The total deficit is the sum of hidden-heterogeneity loss and validation-transfer loss.

Figure 2 shows why the source of validation data matters. Even if the user had to use one common authority weight, the deployment distribution would call for w_P. Using w_Q adds a second loss because the validation distribution is easier than the deployment distribution.

The following table collects the notation used in this section and in the formal model.

Symbol	Meaning
T, t	task type and a realization of task type
F, U	familiar and unfamiliar task types in the two-type benchmark
Q, P, R	validation, deployment, and generic distributions over task types
q, p	unfamiliar-task shares under Q and P in the benchmark
$\theta = \mu_0 + x$	payoff state, baseline mean, and residual state
y	centered AI advice signal for the residual state
A	residual state variance in the scalar benchmark
$v(t), v_F, v_U$	AI error variance by type in the scalar benchmark
$\bar{v}_Q, \bar{v}_P$	average AI error variance under validation and deployment distributions
$b(t)$	covariance between the residual state and the AI signal in the general model
$d(t)$	variance of the AI signal in the general model
$w, w^*(t)$	generic authority weight and type-specific optimal authority weight
w_Q, w_P	validation-induced and deployment-optimal common authority weights
M	information signal or validation partition observed by the user
$W_t(w), W_P(w)$	type-specific and deployment-distribution welfare under authority w
$G_t(w)$	welfare gain from authority w relative to no AI on type t
λ	latent task commonness in the reverse winner's curse model
F_j, K	contributor j's familiarity indicator and aggregate training support
Z, S, X	information environment, AI output, and user information in the Clever Hans bound
c	marginal cost parameter in the quadratic payoff

Table 2: Main notation.

4 A general model of AI authority

The main model is stated at the level of second moments. Normality is not required. The scalar additive benchmark in Section 3 is a special case. Under a Gaussian specialization, the linear authority rule is also the Bayesian posterior mean.

Assumption 1 (Second-moment advice environment)**.** There is a task type $T \in \mathcal{T}$, a residual payoff state x, and a centered AI signal y. For every type t considered in deployment,

$$\mathbb{E}[x \mid T = t] = 0, \qquad \mathbb{E}[y \mid T = t] = 0,$$

and the second moments

$$b(t) := \mathrm{Cov}(x, y \mid T = t), \qquad d(t) := \mathrm{Var}(y \mid T = t)$$

exist with $d(t) > 0$.

The assumption centers the variables after the user's ordinary information and after any observable calibration of the AI signal. The object $b(t)$ measures task-specific covariance between advice and the residual state. The object $d(t)$ measures task-specific signal variance. If task labels also shift the baseline payoff mean, the model applies to the residual state after those mean differences have been accounted for. The general welfare identity permits $b(t)$ to be negative. Statements about positive AI use below impose $b(t) > 0$ when needed; if $b(t) < 0$, the best linear rule reverses the signal.

Assumption 2 (Quadratic payoff and linear authority)**.** The payoff state is $\theta = \mu_0 + x$. The decision maker chooses $a \in \mathbb{R}$ and receives

$$u(a, \theta) = a\theta - \frac{c}{2}a^2, \qquad c > 0.$$

Given an authority weight w, the decision maker forms the forecast

$$\hat{\mu}_w = \mu_0 + wy$$

and chooses

$$a_w = \frac{\hat{\mu}_w}{c}.$$

The quadratic payoff makes welfare comparisons exact. Together with the second-moment assumptions, it delivers the best linear rule below. Under a Gaussian version of the scalar additive benchmark, that rule is also the exact Bayesian rule.

The first formal step is to compute welfare from a fixed authority weight. This calculation matters because the rest of the paper studies how validation evidence selects that weight and what is lost when the selected weight is transferred to a different task distribution.

Lemma 2 (Welfare from a linear authority weight)**.** *For type t, expected welfare from au-*

thority weight w is

$$W_t(w) = \frac{\mu_0^2}{2c} + \frac{wb(t)}{c} - \frac{w^2 d(t)}{2c}.$$

The unique type-specific optimal linear authority weight is

$$w^*(t) = \frac{b(t)}{d(t)},$$

and the type-specific first-best linear value is

$$W_t^* = \frac{\mu_0^2}{2c} + \frac{1}{2c}\frac{b(t)^2}{d(t)}.$$

Proof. Using $a_w = (\mu_0 + wy)/c$ and $\theta = \mu_0 + x$,

$$\mathbb{E}[a_w \theta \mid t] = \frac{1}{c}\mathbb{E}[(\mu_0 + wy)(\mu_0 + x) \mid t] = \frac{1}{c}(\mu_0^2 + wb(t)).$$

The cost term is

$$\mathbb{E}\left[\frac{c}{2}a_w^2 \mid t\right] = \frac{1}{2c}\mathbb{E}[(\mu_0 + wy)^2 \mid t] = \frac{1}{2c}(\mu_0^2 + w^2 d(t)).$$

Subtracting the cost term from the benefit term gives

$$W_t(w) = \frac{\mu_0^2}{2c} + \frac{wb(t)}{c} - \frac{w^2 d(t)}{2c}.$$

This is a strictly concave quadratic in w because $d(t) > 0$. The first-order condition is

$$\frac{b(t)}{c} - \frac{wd(t)}{c} = 0,$$

so $w^*(t) = b(t)/d(t)$. Substitution gives the value formula. □

The lemma identifies the sufficient statistic for linear AI authority. The optimal authority

weight is the covariance between the residual state and the AI signal divided by the variance of the signal. The proof substitutes the linear forecast into the quadratic payoff, expands the expected benefit and cost terms, and obtains a concave quadratic in the authority weight. The first-order condition gives the type-specific weight, and substituting that weight gives the first-best linear value.

5 Validation transfer

Let Q denote the validation distribution over task types and let P denote the deployment distribution. For any distribution R over task types, define

$$b_R := \mathbb{E}_R[b(T)], \qquad d_R := \mathbb{E}_R[d(T)], \qquad w_R := \frac{b_R}{d_R}.$$

The weight w_R is the best common linear authority weight for tasks drawn from R. A user who validates on Q and deploys on P applies w_Q to tasks drawn from P.

The comparison uses three authority rules. The type-specific rule is $w^*(t) = b(t)/d(t)$. The best common rule for deployment is w_P. The transferred validation rule is w_Q. Define

$$W_P^{FB} = \frac{\mu_0^2}{2c} + \frac{1}{2c}\mathbb{E}_P\left[\frac{b(T)^2}{d(T)}\right]$$

as the type-conditional first-best welfare under P. Define

$$W_P(w) = \frac{\mu_0^2}{2c} + \frac{wb_P}{c} - \frac{w^2 d_P}{2c}$$

as welfare under a common authority weight w in deployment.

The main result asks what is lost when an authority weight learned from validation is used in deployment. The answer separates two conceptually different mistakes: treating heterogeneous deployment tasks as if they called for one weight, and using a common weight

learned from the wrong task distribution.

Theorem 2 (Validation-transfer decomposition)**.** *Assume* $\mathbb{E}_P[b(T)^2/d(T)] < \infty$ *and* $d_P, d_Q > 0$. *Then*

$$W_P^{FB} - W_P(w_Q) = \frac{1}{2c}\left[\mathbb{E}_P\left(\frac{b(T)^2}{d(T)}\right) - \frac{b_P^2}{d_P}\right] + \frac{d_P}{2c}(w_Q - w_P)^2.$$

Both terms are weakly positive. The first term is zero if and only if $b(T)/d(T)$ *is* P*-almost surely constant. The second term is zero if and only if* $w_Q = w_P$.

Proof. Taking expectations in Lemma 2 under deployment distribution P gives

$$W_P(w) = \frac{\mu_0^2}{2c} + \frac{wb_P}{c} - \frac{w^2 d_P}{2c}.$$

Maximizing this quadratic over a common weight gives $w_P = b_P/d_P$. The value of the best common deployment weight is

$$W_P(w_P) = \frac{\mu_0^2}{2c} + \frac{1}{2c}\frac{b_P^2}{d_P}.$$

The type-conditional first best uses $w^*(T) = b(T)/d(T)$ on each type, so

$$W_P^{FB} = \frac{\mu_0^2}{2c} + \frac{1}{2c}\mathbb{E}_P\left[\frac{b(T)^2}{d(T)}\right].$$

Add and subtract $W_P(w_P)$:

$$W_P^{FB} - W_P(w_Q) = (W_P^{FB} - W_P(w_P)) + (W_P(w_P) - W_P(w_Q)).$$

The first bracket is

$$\frac{1}{2c}\left[\mathbb{E}_P\left(\frac{b(T)^2}{d(T)}\right) - \frac{b_P^2}{d_P}\right].$$

The second bracket is the loss from using w_Q instead of the maximizer w_P in the quadratic

$W_P(w)$:

$$W_P(w_P) - W_P(w_Q) = \frac{d_P}{2c}(w_Q - w_P)^2.$$

The square is nonnegative. The first term is also nonnegative because the function $(b, d) \mapsto b^2/d$ is convex on $d > 0$, and Jensen's inequality gives

$$\mathbb{E}_P\left[\frac{b(T)^2}{d(T)}\right] \geq \frac{\mathbb{E}_P[b(T)]^2}{\mathbb{E}_P[d(T)]} = \frac{b_P^2}{d_P}.$$

The equality condition in the first term follows from the equality condition in Cauchy–Schwarz: equality holds if and only if $b(T)/d(T)$ is P-almost surely constant. The equality condition in the second term follows immediately from the squared-distance term. □

The theorem is the central decomposition. The first term is the within-deployment loss from using one authority weight even though different task types call for different weights. The second term is the cross-distribution loss from learning the common weight under the validation distribution rather than the deployment distribution. The proof first computes the best common weight for deployment, then adds and subtracts its value. This separates the loss from using one common weight from the extra loss caused by transferring the wrong common weight. Convexity of the map $(b, d) \mapsto b^2/d$ gives nonnegativity of the first term, and the quadratic welfare formula gives the squared-distance form of the second term.

The decomposition measures welfare loss, but it is also useful to know the direction of the authority error on each task type. The next corollary records the simple comparison that determines whether validation transfer causes overreliance or underreliance.

Corollary 1 (Bidirectional misreliance)**.** *Under validation weight w_Q, the decision maker overweights AI advice on type t relative to the type-specific optimum if and only if*

$$w_Q > \frac{b(t)}{d(t)},$$

and underweights AI advice if and only if

$$w_Q < \frac{b(t)}{d(t)}.$$

Proof. The type-specific optimal weight is $b(t)/d(t)$ by Lemma 2. The validation-induced weight is w_Q. Comparing the two weights gives the two inequalities in the statement. □

The corollary clarifies that validation transfer can create both overreliance and underreliance. A pooled validation weight is not simply too high or too low in general; it is compared type by type with the authority ratio appropriate for that type. The proof uses the optimal authority formula from Lemma 2 and compares the transferred weight directly with the type-specific optimum.

5.1 The scalar additive benchmark

The two-type benchmark is a special case of the following scalar additive model. Let

$$y = x + \varepsilon_T,$$

where

$$\mathrm{Var}(x \mid T = t) = A > 0, \qquad \mathbb{E}[\varepsilon_T \mid T = t] = 0,$$

$$\mathrm{Cov}(x, \varepsilon_T \mid T = t) = 0, \qquad \mathrm{Var}(\varepsilon_T \mid T = t) = v(t) > 0.$$

Then

$$b(t) = A, \qquad d(t) = A + v(t), \qquad w^*(t) = \frac{A}{A + v(t)}.$$

This specialization turns the authority ratio into a simple signal-to-noise expression.

The general theorem is easiest to interpret in the scalar additive benchmark. In that case, task heterogeneity appears only through the error variance of the AI signal, so the validation-transfer loss can be written directly in terms of average validation and deployment

error variances.

Corollary 2 (Scalar validation transfer)**.** *In the scalar additive benchmark, define*

$$\bar{v}_R := \mathbb{E}_R[v(T)], \qquad R \in \{P, Q\}.$$

A validation sample from Q induces

$$w_Q = \frac{A}{A + \bar{v}_Q}.$$

The first-best deficit from deploying this weight under P is

$$W_P^{FB} - W_P(w_Q) = \frac{A^2}{2c}\left\{\mathbb{E}_P\left[\frac{1}{A + v(T)}\right] - \frac{1}{A + \bar{v}_P}\right\} + \frac{A + \bar{v}_P}{2c}\left(\frac{A}{A + \bar{v}_Q} - \frac{A}{A + \bar{v}_P}\right)^2.$$

If $P = Q$, the second term is zero and the loss is

$$\frac{A^2}{2c}\left\{\mathbb{E}_P\left[\frac{1}{A + v(T)}\right] - \frac{1}{A + \mathbb{E}_P[v(T)]}\right\}.$$

The decision maker overweights AI on type t if and only if $v(t) > \bar{v}_Q$, and underweights it if and only if $v(t) < \bar{v}_Q$.

Proof. In the scalar additive benchmark,

$$b(t) = \mathrm{Cov}(x, x + \varepsilon_T \mid t) = \mathrm{Var}(x \mid t) = A,$$

and

$$d(t) = \mathrm{Var}(x + \varepsilon_T \mid t) = A + v(t).$$

Therefore $w^*(t) = A/(A + v(t))$. For any distribution R,

$$b_R = A, \qquad d_R = A + \bar{v}_R, \qquad w_R = \frac{A}{A + \bar{v}_R}.$$

Substituting these expressions into Theorem 2 gives the stated decomposition. Since $A/(A+v)$ is strictly decreasing in v, $w_Q > w^*(t)$ is equivalent to $v(t) > \bar{v}_Q$, and $w_Q < w^*(t)$ is equivalent to $v(t) < \bar{v}_Q$. □

The scalar formula is the simplest version of validation transfer. If validation is easier than deployment, then $\bar{v}_Q < \bar{v}_P$ and the transferred weight is too high as a common deployment weight. If validation and deployment coincide, the only remaining loss is the Jensen gap from hiding heterogeneity.

Recall the convex order. A random variable V_2 is more dispersed than V_1 in the convex order if $\mathbb{E}[V_1] = \mathbb{E}[V_2]$ and $\mathbb{E}[\phi(V_2)] \geq \mathbb{E}[\phi(V_1)]$ for every convex function ϕ for which the expectations exist.

The scalar formula makes clear that the mean error variance is not enough. The next result shows that, even when pooled validation estimates the same average reliability, greater hidden dispersion in task-level reliability increases the value left on the table by a pooled authority rule.

Proposition 1 (Hidden heterogeneity)**.** *In the scalar additive benchmark with $P = Q$, pooled welfare depends only on $\bar{v} = \mathbb{E}[v(T)]$. If two error-variance distributions have the same mean and one is more dispersed in the convex order, then the more dispersed distribution has weakly larger first-best deficit under pooled validation. If $v(T) = \bar{v} + \eta(T)$, $\mathbb{E}[\eta(T)] = 0$, and the heterogeneity is small, then*

$$W^{FB} - W^{pool} = \frac{A^2}{2c} \frac{\mathrm{Var}(\eta)}{(A+\bar{v})^3} + o\left(\mathbb{E}[\eta^2]\right).$$

Proof. When $P = Q$, the pooled weight is $A/(A+\bar{v})$. Pooled welfare is

$$W^{pool} = \frac{\mu_0^2}{2c} + \frac{A^2}{2c(A+\bar{v})},$$

which depends on the distribution of $v(T)$ only through its mean. The first-best value is

$$W^{FB} = \frac{\mu_0^2}{2c} + \frac{A^2}{2c}\mathbb{E}\left[\frac{1}{A + v(T)}\right].$$

The function $f(v) = 1/(A + v)$ is convex on $v > 0$. Therefore a mean-preserving increase in dispersion raises $\mathbb{E}[f(v(T))]$ and weakly raises $W^{FB} - W^{pool}$.

For the local approximation, let $v(T) = \bar{v} + \eta(T)$ with $\mathbb{E}[\eta(T)] = 0$. A second-order Taylor expansion around $\bar{v}$ gives

$$\mathbb{E}[f(\bar{v} + \eta)] - f(\bar{v}) = \frac{1}{2}f''(\bar{v})\,\mathrm{Var}(\eta) + o(\mathbb{E}[\eta^2]).$$

Since $f''(\bar{v}) = 2/(A + \bar{v})^3$, multiplying by $A^2/(2c)$ gives the stated expression. □

The proposition shows why an average validation score is least informative when reliability is uneven across tasks. Holding the mean error variance fixed, greater dispersion raises the gap between conditional authority and pooled authority. The proof observes that pooled welfare depends only on the mean error variance, while first-best welfare depends on the expectation of a convex function of the error variance. Convex-order dispersion therefore increases the first-best value without changing the pooled value. The local expansion is the second-order Taylor approximation of that same Jensen gap.

6 Conditional validation

It is useful to think of conditional validation as a table that partitions validation evidence by observable task features. Each cell of the table corresponds to one conditioning event, such as a particular familiarity category, task domain, or difficulty class, and reports how reliable the AI signal is within that subset of tasks.

A validation report is useful when it tells the user which reliability cell a task belongs to. Let M be a signal about task type. Conditional on message $M = m$, define the cell-level

authority weight

$$w(m) = \frac{\mathbb{E}[b(T) \mid M = m]}{\mathbb{E}[d(T) \mid M = m]}.$$

A signal M' Blackwell-dominates M if M can be generated as a garbling of M'.

The preceding section shows why pooling can be costly. The natural response is to ask what kind of additional validation information is valuable. The next theorem answers this question using the Blackwell order: a validation signal is valuable when it refines the information used to choose the authority weight.

Theorem 3 (Blackwell value of conditional validation)**.** *For any signal M about task type,*

$$W(M) = \frac{\mu_0^2}{2c} + \frac{1}{2c}\mathbb{E}\left[\frac{\mathbb{E}[b(T) \mid M]^2}{\mathbb{E}[d(T) \mid M]}\right].$$

If M' Blackwell-dominates M, then

$$W(M') \geq W(M).$$

Proof. Given message $M = m$, the conditional welfare from a cell-level weight w is

$$\frac{\mu_0^2}{2c} + \frac{w\mathbb{E}[b(T) \mid M = m]}{c} - \frac{w^2\mathbb{E}[d(T) \mid M = m]}{2c}.$$

The maximizing cell-level weight is

$$w(m) = \frac{\mathbb{E}[b(T) \mid M = m]}{\mathbb{E}[d(T) \mid M = m]}.$$

Substituting this weight and taking expectations over M gives the formula for $W(M)$.

Let

$$\phi(B, D) = \frac{B^2}{D}, \qquad D > 0.$$

The function ϕ is convex. If M' Blackwell-dominates M, then the conditional distribution

of $(\mathbb{E}[b(T) \mid M'], \mathbb{E}[d(T) \mid M'])$ given M has mean $(\mathbb{E}[b(T) \mid M], \mathbb{E}[d(T) \mid M])$. Jensen's inequality gives

$$\mathbb{E}\left[\phi(\mathbb{E}[b(T) \mid M'], \mathbb{E}[d(T) \mid M']) \mid M\right] \geq \phi(\mathbb{E}[b(T) \mid M], \mathbb{E}[d(T) \mid M]).$$

Taking expectations gives $W(M') \geq W(M)$. Perfect type information gives $\mathbb{E}[b(T) \mid T] = b(T)$ and $\mathbb{E}[d(T) \mid T] = d(T)$. An uninformative signal gives $\mathbb{E}[b(T)]$ and $\mathbb{E}[d(T)]$. □

The theorem formalizes the value of conditional validation. A more informative validation signal lets the user choose authority weights on finer cells of the task distribution. The proof first solves the cell-level quadratic problem. It then compares two validation signals using Jensen's inequality applied to the convex function $(B, D) \mapsto B^2/D$. A Blackwell refinement preserves conditional means at the coarser level while revealing more information before the authority rule is chosen, so ex ante welfare cannot fall. Appendix B shows that this is an ex ante result and need not imply higher welfare for every realized task type.

In the scalar additive benchmark, Theorem 3 becomes

$$W(M) = \frac{\mu_0^2}{2c} + \frac{A^2}{2c}\mathbb{E}\left[\frac{1}{A + \mathbb{E}[v(T) \mid M]}\right].$$

The relevant report is not a verbal explanation by itself. The relevant report is information that changes the conditional error variance used to choose authority.

7 Learning from pooled validation

Suppose a labeled validation sample is drawn from distribution Q. In observation i, the centered state is x_i and the centered AI signal is y_i. A one-weight learner estimates w by least squares:

$$\hat{w}_n \in \arg\min_w \sum_{i=1}^{n} (x_i - w y_i)^2.$$

When $\sum_i y_i^2 > 0$,

$$\hat{w}_n = \frac{\sum_i x_i y_i}{\sum_i y_i^2}.$$

The next result shows that validation transfer is not a small-sample problem. Even with a large validation sample, pooled learning converges to the best pooled projection under the validation distribution. That projection can still be the wrong decision object for heterogeneous deployment.

Proposition 2 (Pooled projection learned from validation)**.** *Assume validation observations are stationary ergodic with limiting distribution Q and finite second moments, and suppose $d_Q > 0$. Then*

$$\hat{w}_n \to \frac{b_Q}{d_Q} = w_Q.$$

If the observations are iid and have finite fourth moments, then $\sqrt{n}(\hat{w}_n - w_Q)$ converges in distribution to a centered normal random variable.

Proof. The least-squares estimator can be written as

$$\hat{w}_n = \frac{n^{-1}\sum_{i=1}^n x_i y_i}{n^{-1}\sum_{i=1}^n y_i^2}.$$

Stationary ergodicity and finite second moments imply

$$n^{-1}\sum_{i=1}^n x_i y_i \to \mathbb{E}_Q[xy] = b_Q, \qquad n^{-1}\sum_{i=1}^n y_i^2 \to \mathbb{E}_Q[y^2] = d_Q.$$

Since $d_Q > 0$, the continuous mapping theorem gives $\hat{w}_n \to b_Q/d_Q$.

Under iid sampling with finite fourth moments, the vector

$$\left(n^{-1}\sum_i x_i y_i,\ n^{-1}\sum_i y_i^2\right)$$

is asymptotically normal by the multivariate central limit theorem. Applying the delta method to the ratio map $(u, v) \mapsto u/v$ gives asymptotic normality of $\hat{w}_n$ at rate $\sqrt{n}$. □

The proposition separates statistical consistency from decision relevance. A user can consistently learn the pooled validation projection and still transfer the wrong authority rule to deployment. The proof writes the least-squares estimate as a ratio of sample moments. Ergodic convergence sends the numerator and denominator to their validation-distribution expectations, and the continuous mapping theorem gives the probability limit. Under iid sampling and fourth moments, the central limit theorem and delta method deliver asymptotic normality.

Appendix A gives a local detection result for omitted variance heterogeneity in the scalar Gaussian case. The result shows that learning an average variance and detecting hidden heterogeneity are different statistical tasks.

8 Use, consultation, and the no-AI benchmark

Overreliance relative to the type-specific first best does not imply harm relative to no AI. This distinction matters for policy. A user can put too much authority on AI advice and still be better off than without the advice.

The welfare effect of excessive authority has two thresholds. A weight can exceed the first-best type-specific authority weight and yet still be better than ignoring the AI signal. The next result identifies when overuse crosses the stronger threshold into harm relative to no AI.

Proposition 3 (Harm relative to no AI)**.** *For type t, the welfare gain from applying authority weight w relative to using no AI advice is*

$$G_t(w) = W_t(w) - W_t(0) = \frac{wb(t)}{c} - \frac{w^2 d(t)}{2c}.$$

If $b(t) > 0$ and $w > 0$, then

$$G_t(w) = \frac{d(t)}{2c} w \left(2w^*(t) - w\right),$$

so

$$G_t(w) < 0 \quad \Longleftrightarrow \quad w > 2w^*(t).$$

At the aggregate deployment level, if $w_Q > 0$ *and* $b_P > 0$*, then*

$$W_P(w_Q) < W_P(0) \quad \Longleftrightarrow \quad w_Q > 2w_P.$$

Proof. The gain from authority w relative to no AI is obtained by subtracting $W_t(0)$ from the welfare formula in Lemma 2:

$$G_t(w) = \frac{wb(t)}{c} - \frac{w^2 d(t)}{2c}.$$

Since $w^*(t) = b(t)/d(t)$,

$$G_t(w) = \frac{d(t)}{2c} w(2w^*(t) - w).$$

When $b(t) > 0$ and $w > 0$, the sign is the sign of $2w^*(t) - w$. This gives the type-level condition.

For the aggregate condition, replace $b(t)$ and $d(t)$ by b_P and d_P. The deployment-optimal common weight is $w_P = b_P/d_P$. The same argument gives $W_P(w_Q) < W_P(0)$ if and only if $w_Q > 2w_P$. □

The proposition distinguishes using too much AI from being worse off than using no AI. Moderate overreliance can still improve welfare if the advice remains informative enough. The proof subtracts no-AI welfare from the type-level welfare formula and factors the gain into a term proportional to $w(2w^*(t) - w)$. This shows that harm arises only when authority exceeds twice the type-specific optimum. The aggregate statement repeats the same calculation after replacing type-level moments by deployment averages.

In the scalar additive benchmark, transferring $w_Q = A/(A + \bar{v}_Q)$ to distribution P is

harmful relative to no AI if and only if

$$\bar{v}_P > A + 2\bar{v}_Q.$$

Deployment must be much noisier than validation before the transferred advice is worse than no advice at the aggregate level.

The same scalar benchmark also gives a selection point. Suppose the user can choose whether to consult the AI at cost k. Let the user's residual uncertainty be A_t on type t, and suppose the user evaluates consultation using the pooled validation variance $\bar{v}_Q$. The perceived value of consultation on a task with residual variance A is

$$\Delta(A; \bar{v}_Q) = \frac{A^2}{2c(A + \bar{v}_Q)}.$$

This is the perceived gain from adding the AI signal when the user treats its error variance as $\bar{v}_Q$.

The final deployment result studies selection into AI use. Users may be especially likely to seek AI help when their own residual uncertainty is high. That behavior is sensible under the validation report they observe, but it can direct AI use toward the exact tasks where the report transfers poorly.

Proposition 4 (Selected consultation in the scalar benchmark)**.** *Let $A_F, A_U > 0$ be type-specific residual variances, and suppose $A_U > A_F$. For fixed $\bar{v}_Q > 0$, $\Delta(A; \bar{v}_Q)$ is strictly increasing in $A > 0$. For every consultation cost*

$$k \in (\Delta(A_F; \bar{v}_Q), \Delta(A_U; \bar{v}_Q)),$$

the user consults AI on unfamiliar tasks and not on familiar tasks. If the actual gain from consultation on unfamiliar tasks is negative, this selected consultation occurs only on a type where AI use lowers welfare relative to no AI.

Proof. For fixed $\bar{v}_Q > 0$,

$$\Delta(A; \bar{v}_Q) = \frac{A^2}{2c(A + \bar{v}_Q)}.$$

Differentiating with respect to A gives

$$\frac{\partial \Delta}{\partial A} = \frac{A(A + 2\bar{v}_Q)}{2c(A + \bar{v}_Q)^2} > 0.$$

Thus $\Delta(A_U; \bar{v}_Q) > \Delta(A_F; \bar{v}_Q)$ whenever $A_U > A_F$. Any consultation cost

$$k \in (\Delta(A_F; \bar{v}_Q), \Delta(A_U; \bar{v}_Q))$$

then lies above the familiar-task perceived consultation value and below the unfamiliar-task perceived consultation value. The user therefore consults on unfamiliar tasks and not on familiar tasks. If the true unfamiliar-task gain is negative, this selected consultation occurs only where advice lowers welfare relative to no AI. □

The proposition explains why AI may be consulted precisely on the tasks for which validation evidence is least portable. Higher residual uncertainty raises the perceived value of advice under the pooled validation rule. The proof differentiates the perceived gain with respect to residual variance and shows that it is increasing. A consultation cost can therefore lie between the perceived value for familiar and unfamiliar tasks, inducing consultation only on unfamiliar tasks. If the true gain there is negative, selection into AI use is welfare reducing.

9 Conclusion

AI is often described as the ultimate scaling technology. It allows organizations to extend scarce expertise, accelerate analysis, and make decisions at a speed that would have been impossible when every task required direct human attention. This is especially important for young firms. A start-up can now perform many functions that previously required a

much larger organization: drafting documents, writing code, analyzing customers, designing experiments, preparing investor materials, and making operational decisions. In this sense, AI appears to relax one of the central constraints on growth. It helps a small team act as if it had access to a much larger stock of human knowledge.

The results in this paper identify an important caution. There are two distinct ways in which a decision maker can lack knowledge. In the first case, the decision maker lacks local knowledge, but the relevant knowledge exists elsewhere. The problem is unfamiliar to the user, but it is familiar in the broader environment. There are examples, precedents, training data, experts, and validation opportunities. In this case, AI can be extremely valuable. It closes a local knowledge gap by drawing on knowledge that exists globally. Many standard business decisions have this form. A manager may not know the best way to write a sales email, structure a hiring rubric, interpret a familiar financial ratio, or summarize a standard legal clause, but the relevant knowledge is widely available. AI can be reliable because the task is unfamiliar locally but common globally.

The second case is different. The decision maker lacks knowledge because the relevant knowledge does not yet exist globally, or exists only weakly. The task is not merely unfamiliar to the user. It is unfamiliar in the broader information environment from which the AI system draws. This is more likely in innovative settings: new markets, new technologies, novel organizational designs, unusual strategic choices, and start-up problems for which there are few close precedents. These are exactly the situations in which AI may be most tempting. The team faces uncertainty, expertise is scarce, and the pressure to move quickly is high. If high uncertainty is correlated with weak global support, these are also the situations in which AI authority can be most dangerous. The system may appear to solve the knowledge gap precisely when the information needed to solve the problem was not present in the data environment that trained or validated it.

This distinction is consistent with the paper's assumption that local familiarity makes global familiarity more likely. The model does not rule out the favorable case in which a

task is locally unfamiliar but globally familiar. Instead, it says that local unfamiliarity shifts posterior beliefs toward weaker global support. AI authority should therefore be limited unless there is independent evidence that the task is well represented in the broader information environment.

The paper's central message is not that AI should be used less. It is that AI authority should be conditioned on the kind of knowledge gap being addressed. If the user lacks knowledge that exists elsewhere, AI can scale expertise. If the world itself lacks the relevant knowledge, AI may scale confidence without scaling understanding. Validation evidence is informative only relative to the distribution of tasks on which it was generated. A pooled validation number can be correct and still be the wrong guide for a particular decision if the deployment task differs from the validation environment.

The winner's curse intuition gives a practical rule. In a common-value auction, a bidder who wins should shade his bid because winning is bad news about the information held by others. The more severe the selection problem, the more the bidder should shade. A similar logic applies to AI reliance. A user's lack of familiarity is not merely a private fact about the user. It may be evidence that the task is also weakly represented in the broader information environment. The user should therefore shade trust in AI when unfamiliarity is informative about low global support. Counterintuitively, this means that the human work component may need to increase in some of the cases where AI seems most useful. When AI appears to answer the greatest need, the need itself may be evidence that the AI has weaker support.

This distinction is especially important for innovative firms. Established businesses often face repeated problems embedded in rich bodies of prior knowledge. Start-ups often face problems whose novelty is the source of their value. AI can help start-ups move faster, but speed is not the same as validation. The more a problem lies outside established knowledge, the more the organization should treat AI output as a hypothesis to be tested rather than an authority to be followed. The right organizational response is not to reject AI, but to pair AI use with more deliberate human verification, experimentation, and domain-specific

judgment where the task is novel.

The broader implication is that AI changes the allocation of human effort rather than eliminating the need for it. For familiar and well-supported tasks, human effort can often move away from execution and toward supervision. For unfamiliar and weakly supported tasks, human effort must move toward validation, experimentation, and the creation of new knowledge. The danger is to reverse this allocation by trusting AI most where the user knows least. The paper shows why that instinct can be systematically wrong. The right lesson is to scale AI where knowledge is globally available, and to shade AI authority where the absence of local knowledge may be evidence of an absence of global knowledge.

A Local detection lower bound

Fix $\pi \in (0,1)$ and $\bar{v} > 0$. Let

$$Q = \mathcal{N}(0, \bar{v})$$

and consider local two-type alternatives

$$P_h = \pi \mathcal{N}\left(0, \bar{v}[1-(1-\pi)h]\right) + (1-\pi)\mathcal{N}\left(0, \bar{v}[1+\pi h]\right),$$

for $h > 0$ small enough that both component variances are positive.

Proposition 5 (Local lower bound for detecting omitted variance heterogeneity)**.** *Under the alternatives P_h, the variance equals $\bar{v}$ and the excess kurtosis is*

$$\kappa(h) = 3\pi(1-\pi)h^2.$$

Moreover,

$$\chi^2(P_h \| Q) = \frac{\kappa(h)^2}{24} + O(h^5).$$

For n independent observations, any level-α test of Q against P_h has power at most

$$\alpha + \sqrt{\frac{(1+\chi^2(P_h\|Q))^n - 1}{2}}.$$

For any target power $\beta > \alpha$, achieving power at least β requires

$$n \geq \frac{\log(1+2(\beta-\alpha)^2)}{\log(1+\chi^2(P_h\|Q))},$$

and locally $n = \Omega(\kappa(h)^{-2})$.

Proof. Let $Y = X/\sqrt{\bar{v}}$, so under Q, $Y \sim \mathcal{N}(0,1)$. For a Gaussian component with relative

variance $1 + a$, its likelihood ratio relative to Q is

$$r_a(Y) = (1 + a)^{-1/2} \exp\left(\frac{Y^2}{2}\frac{a}{1 + a}\right).$$

For small a,

$$r_a(Y) = 1 + \frac{a}{2}H_2(Y) + \frac{a^2}{8}H_4(Y) + O(a^3),$$

where $H_2(Y) = Y^2 - 1$ and $H_4(Y) = Y^4 - 6Y^2 + 3$.

In P_h, the two relative variance perturbations are

$$a_F = -(1 - \pi)h, \qquad a_U = \pi h.$$

The mixture average is zero and the mixture second moment is $\pi(1 - \pi)h^2$. The likelihood ratio $L_h = dP_h/dQ$ therefore satisfies

$$L_h(Y) - 1 = \frac{\pi(1 - \pi)h^2}{8}H_4(Y) + O(h^3).$$

Since $\mathbb{E}_Q[H_4(Y)^2] = 24$,

$$\chi^2(P_h\|Q) = \mathbb{E}_Q[(L_h - 1)^2] = \frac{3}{8}\pi^2(1 - \pi)^2h^4 + O(h^5).$$

Because $\kappa(h)^2/24 = (3/8)\pi^2(1 - \pi)^2h^4$, the chi-square expansion follows.

For product measures,

$$\chi^2(P_h^n\|Q^n) = (1 + \chi^2(P_h\|Q))^n - 1.$$

Also,

$$\mathrm{TV}(P_h^n, Q^n) \le \sqrt{\frac{\chi^2(P_h^n\|Q^n)}{2}}.$$

If φ is a level-α test, then

$$\mathbb{E}_{P_h^n}[\varphi] \leq \mathbb{E}_{Q^n}[\varphi] + \mathrm{TV}(P_h^n, Q^n) \leq \alpha + \sqrt{\frac{(1+\chi^2(P_h\|Q))^n - 1}{2}}.$$

Solving this upper bound for a target power β gives the sample-size condition. The local rate follows from the chi-square expansion. □

This shows that pooled variance can be easy to estimate while omitted variance heterogeneity remains statistically hard to detect. The proof constructs local mixture alternatives with the same variance as the benchmark normal distribution. Because the variance is matched, the first detectable difference appears through higher moments. The chi-square expansion then bounds the total variation distance between the null and local alternatives, which gives the stated upper bound on test power.

B Failure of typewise Pareto monotonicity

The Blackwell theorem is an ex ante statement. It does not imply that every true task type is better off after every refinement of the validation signal. An example illustrates this limitation.

Proposition 6 (Typewise Pareto monotonicity fails)**.** *There exist scalar-additive parameters and signal accuracies $\delta' > \delta$ such that increasing the accuracy of a binary task-type signal lowers welfare conditional on the true unfamiliar type U.*

Proof. Set

$$A = 1, \qquad c = 1, \qquad \mu_0 = 1, \qquad \pi = \frac{4}{5}, \qquad v_F = \frac{1}{10}, \qquad v_U = 10.$$

At $\delta = 1/2$, the type signal is uninformative. The pooled variance is

$$\bar{v} = \frac{4}{5} \cdot \frac{1}{10} + \frac{1}{5} \cdot 10 = \frac{52}{25},$$

so

$$\hat{w} = \frac{1}{1+\bar{v}} = \frac{25}{77}.$$

Conditional on the true type U,

$$W_U(1/2) = \frac{1}{2} + \frac{25}{77} - \frac{11}{2}\left(\frac{25}{77}\right)^2 = \frac{12}{49}.$$

Now set $\delta' = 3/5$. If the message is U, Bayes' rule gives

$$\mathbb{P}(F \mid M = U) = \frac{8}{11},$$

so

$$\mathbb{E}[v(T) \mid M = U] = \frac{14}{5}, \qquad w(U) = \frac{5}{19}.$$

If the message is F, Bayes' rule gives

$$\mathbb{P}(F \mid M = F) = \frac{6}{7},$$

so

$$\mathbb{E}[v(T) \mid M = F] = \frac{53}{35}, \qquad w(F) = \frac{35}{88}.$$

Conditional on the true type U, the correct message U arrives with probability $3/5$, and the mistaken message F arrives with probability $2/5$. Hence

$$W_U(3/5) = \frac{1}{2} + \frac{3}{5}\left[\frac{5}{19} - \frac{11}{2}\left(\frac{5}{19}\right)^2\right] + \frac{2}{5}\left[\frac{35}{88} - \frac{11}{2}\left(\frac{35}{88}\right)^2\right] = \frac{61107}{254144}.$$

Since

$$\frac{61107}{254144} < \frac{12}{49},$$

$W_U(3/5) < W_U(1/2)$. The refinement raises ex ante welfare by the Blackwell theorem, but it lowers welfare conditional on this true type. □

This shows why the Blackwell result should be read as an ex ante statement. More informative validation improves welfare before the task type and message are realized, but it need not raise welfare for every true type after every possible message. The proof gives scalar parameters and compares welfare under an uninformative signal with welfare under a more accurate signal, conditional on the unfamiliar type. The direct calculation shows that the refined signal can lower welfare for that type even though it raises ex ante value.

C Binary familiar-side decomposition

This decomposition is useful for interpreting where aggregate loss appears in the binary scalar benchmark. It is secondary to the general validation-transfer theorem.

Proposition 7 (Familiar-side share of the aggregate deficit)**.** *In the binary scalar additive benchmark with* $0 < v_F < v_U$, $\mathbb{P}(F) = \pi$, *and*

$$\bar{v} = \pi v_F + (1-\pi)v_U,$$

let $\hat{w} = A/(A+\bar{v})$ *and*

$$L_\tau = \frac{A+v_\tau}{2c}(\hat{w} - w^*(v_\tau))^2, \qquad D_F = \pi L_F, \qquad D_U = (1-\pi)L_U.$$

Then

$$\frac{D_F}{D_F + D_U} = \frac{(1-\pi)(A+v_U)}{(1-\pi)(A+v_U) + \pi(A+v_F)}.$$

Consequently,

$$D_F > D_U \quad \Longleftrightarrow \quad \frac{1-\pi}{\pi} > \frac{A+v_F}{A+v_U}.$$

Proof. Let $\Delta = v_U - v_F > 0$. Since

$$\bar{v} - v_F = (1-\pi)\Delta, \qquad v_U - \bar{v} = \pi\Delta,$$

the deviations from the type-specific weights are

$$\hat{w} - w^*(v_F) = -\frac{A(1-\pi)\Delta}{(A+\bar{v})(A+v_F)},$$

and

$$\hat{w} - w^*(v_U) = \frac{A\pi\Delta}{(A+\bar{v})(A+v_U)}.$$

Therefore

$$L_F = \frac{A^2(1-\pi)^2\Delta^2}{2c(A+\bar{v})^2(A+v_F)},$$

and

$$L_U = \frac{A^2\pi^2\Delta^2}{2c(A+\bar{v})^2(A+v_U)}.$$

Multiplying by type shares and canceling common factors gives the stated share. Comparing the share to 1/2 gives the dominance condition. □

This shows that the aggregate welfare deficit from pooling need not be concentrated on unfamiliar tasks. Familiar tasks may account for most of the loss if they are sufficiently common, because even a smaller per-task underweighting loss can dominate in the aggregate. The proof computes the deviations of the pooled weight from each type-specific weight, multiplies by type shares, and cancels common terms to obtain the familiar-side share.

References

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky. 2025. "Designing Human-AI Collaboration: A Sufficient-Statistic Approach." NBER Working Paper No. 33949, June. doi:10.3386/w33949.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33(2): 31–50.

Andrews, Isaiah, and Emily Oster. 2019. "A Simple Approximation for Evaluating External Validity Bias." *Economics Letters* 178: 58–62.

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu. 2025. "The Transfer Performance of Economic Models." Working paper.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89(1): 133–161.

Balakrishnan, Maya, Kris Johnson Ferreira, and Jordan Tong. 2026. "Human-Algorithm Collaboration with Private Information: Naïve Advice-Weighting Behavior and Mitigation." *Management Science* 72(1): 265–284. doi:10.1287/mnsc.2022.03850.

Bates, J. M., and C. W. J. Granger. 1969. "The Combination of Forecasts." *Operational Research Quarterly* 20(4): 451–468.

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. "A Theory of Learning from Different Domains." *Machine Learning* 79: 151–175.

Berk, Robert H. 1966. "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *Annals of Mathematical Statistics* 37(1): 51–58.

Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics* 24(2): 265–272.

Capen, Edward C., Robert V. Clapp, and William M. Campbell. 1971. "Competitive Bidding in High-Risk Situations." *Journal of Petroleum Technology* 23(6): 641–653.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5(4): 559–583.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2026. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." *Organization Science* 37(2): 403–423. doi:10.1287/orsc.2025.21838.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144(1): 114–126.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64(3): 1155–1170.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica* 84(3): 1093–1130.

Esponda, Ignacio, and Demian Pouzo. 2021. "Asymptotic Behavior of Bayesian Learners with Misspecified Models." *Journal of Economic Theory* 195: 105260.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64(12): 86–92.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Bren-

del, Matthias Bethge, and Felix A. Wichmann. 2020. "Shortcut Learning in Deep Neural Networks." *Nature Machine Intelligence* 2: 665–673.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125(1–2): 241–270.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101(6): 2590–2615.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review: Papers and Proceedings* 105(5): 491–495.

Koh, Pang Wei, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. "WILDS: A Benchmark of in-the-Wild Distribution Shifts." In *Proceedings of the 38th International Conference on Machine Learning*, 5637–5664.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." *Nature Communications* 10: 1096.

Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46(1): 50–80.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and others. 2023. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research*.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2026. "Large Language Models: An Applied Econometric Framework." *Annual Review of Economics* 18: 283–316. doi:10.1146/annurev-economics-120925-105620.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72(4): 1221–1246.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1): 57–91.

Milgrom, Paul R., and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50(5): 1089–1122.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2): 87–106.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39(2): 230–253.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability Across Populations." *Statistical Science* 29(4): 579–595.

Pfungst, Oskar. 1911. *Clever Hans (The Horse of Mr. von Osten): A Contribution to Experimental Animal and Human Psychology*. Translated by Carl L. Rahn. New York: Henry Holt.

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, eds. 2009. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." In *Proceedings of the NeurIPS Datasets and Benchmarks Track*.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. "Do ImageNet Classifiers Generalize to ImageNet?" In *Proceedings of the 36th International Conference on Machine Learning*, 5389–5400.

Steyvers, Mark, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. "Bayesian Modeling of Human-AI Complementarity." *Proceedings of the National Academy of Sciences* 119(11): e2111547119.

Thaler, Richard H. 1988. "Anomalies: The Winner's Curse." *Journal of Economic Perspectives* 2(1): 191–202.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas W. Malone. 2024. "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-analysis." *Nature Human Behaviour* 8: 2293–2303.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function." In *Proceedings of the 41st International Conference on Machine Learning*, 48919–48937.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18(6): 3045–3089.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50(1): 1–25.