The Value of Context: Human versus Algorithmic Evaluators

Andrei Iakovlev and Annie Liang*

August 26, 2025

Abstract

Many predictions previously conducted by human experts (e.g., loan assessments and medical diagnoses) can now be automated. How should individuals choose between institutions where predictions are made by algorithms, and institutions where predictions are made by people? We propose a framework to examine one key distinction: Machine learning algorithms consider a fixed, standardized set of covariates for all individuals, whereas human evaluators adapt the choice of covariates to each person. Our framework defines and analyzes the advantage of this customization—the *value of context*—in environments with complex prediction problems. We show that unless the agent has prior reason to think that their individualized context is especially informative, then the benefit of more information generally outweighs the value of customization.

1 Introduction

"A statistical formula may be highly successful in predicting whether or not a person will go to a movie in the next week. But someone who knows that this person is laid up with a broken leg will beat the formula. No formula can take into account the infinite range of such exceptional events." — Atul Gawande, *Complications: A Surgeon's Notes on an Imperfect Science*

*Department of Economics, Northwestern University. We thank Modibo Camara, Krishna Dasaratha, Alex Frankel, Drew Fudenberg, Ben Golub, Kevin He, Xiaosheng Mu, Matthew Murphy, Jacopo Perego, Debraj Ray, and Marzena Rostek for helpful comments and suggestions.

As algorithmic predictions become increasingly viable alternatives to human judgment, institutions are differentiating into two models: "brick-and-mortar" establishments that emphasize individualized human assessment, and digital platforms that rely on standardized algorithmic predictions. For example, in wealth management, robo-advisors such as Wealthfront automatically allocate and rebalance investment portfolios, while human financial advisors provide more tailored guidance; in consumer lending, fintech algorithms such as Upstart and MYbank automate loan decisions based on large datasets, while relationship-based lenders make case-by-case decisions; in insurance, online platforms such as Lemonade and Insurify use machine learning to automate insurance claims, while traditional insurers rely on human insurance agents; and in medicine, telemedicine and digital health platforms automate the provision of care at scale, while doctors provide specialized attention and care.

One important distinction between algorithmic and human evaluation is that algorithms base their predictions on a standardized set of inputs, while human evaluators can adapt what they learn to each person. For example, an AI lending model might flag unemployment as grounds for denying a loan, overlooking the crucial detail that the applicant left work to launch a startup. If that context is not among the algorithm's predefined inputs, the applicant cannot supply it—but he can inform a human. The perception that humans are better able to account for individuals' unique circumstances is a significant factor in resistance to AI (Longoni et al., 2019). Our objective in this paper is to understand when, and to what extent, this difference between human and machine evaluation matters.

We propose a theoretical framework to compare evaluation based on a large and standardized set of inputs versus evaluation based on a small set of targeted inputs. In our model, an agent chooses between an algorithmic institution and a human institution before knowing what specific prediction problem will be relevant. To fix ideas, consider a small business owner who chooses between a *fintech bank* and a *traditional bank*, where at the time of contracting the agent cannot foresee which financing tools he will need in the future.[1]

Formally, a *prediction problem* consists of a set of types to be predicted—such as the probability with which the agent would pay back a specific loan—and a function relating

[1] At a high level, this question is similar to Akbarpour et al. (2024)'s comparison of the network diffusion value of a small number of targeted seeds versus a larger number of randomly selected seeds. Like them, we will find that a larger number of (non-targeted) inputs is superior, but the mechanisms behind these results are very different; in particular, network structure does not play a role in our results.

agent covariates (i.e., everything that is relevant about the agent's financial situation) to the unknown type. To predict this type, each bank observes some of the agent's covariates, and predicts the average type among individuals who share the agent's observed covariates.

The banks differ in which covariates are observed. Both banks observe the agent's "standard" covariates, such as their credit score and employment status. But there is additionally a large set of "nonstandard" covariates, whose relevance and meaning depends on the realized prediction problem. For example, collateral ownership may be decisive for an equipment loan—since the bank can secure the loan against the collateral—but irrelevant for unsecured credit. And seasonal cash flows may strongly affect the prospects of a short-term bridge loan but carry little weight in a long-term expansion loan. At the time of choosing a bank, the agent is uninformed about which kind of loan he will eventually need, so he does not know which (if any) nonstandard covariates will represent important context. Formally, we assume that before the prediction problem is realized, the agent's type is distributed independently of his nonstandard covariates.

Of the nonstandard covariates, the fintech bank observes a large fraction that is standardized across individuals, while the traditional bank observes a smaller fraction that is personalized to each individual. To compare the two banks, we adopt a conservative criterion. We say that the agent robustly prefers the fintech bank if his expected payoff is higher there than at the traditional bank, even when the traditional bank observes the nonstandard covariates most favorable to the agent (e.g., indicating the highest probability of repayment when the agent's payoffs are increasing in this prediction). And we instead say that the agent robustly prefers the traditional bank if his expected payoff is higher there than at the fintech bank, even when the traditional bank observes the nonstandard covariates least favorable to him.

A central parameter in our analysis is the total number of nonstandard covariates. Because the agent's type is fully determined by the complete covariate vector, this number does not capture the total amount of information but instead reflects the potential complexity of the prediction problem. We refer to settings where the outcome may depend on many covariates as *richer*.

Our first main result says that as the covariate space becomes richer, the benefit to targeted covariate selection vanishes. Formally, we use as a benchmark the agent's expected payoff when the prediction is based only on his standard covariates. We then define the *value*

of context as the best-case improvement in payoffs when some fixed fraction of covariates are acquired, relative to this benchmark. That is, the value of context is the largest possible utility gain when the agent's "best" nonstandard covariates are observed.

We prove that the expected value of context converges to zero as the covariate space grows large. Thus even though there may be realizations of the prediction problem given which the value of context is large, in *expectation* it is not. The contrapositive of this result is that if the expected value of context is high, it must be that the agent has some ex-ante knowledge about the predictive roles of the nonstandard covariates.

We prove this result by studying the sensitivity of the evaluator's expectation to the set of covariates that are revealed. Intuitively, a large value of context requires that the evaluator's beliefs move sharply after observing certain nonstandard covariates. We show that the largest feasible change in the evaluator's beliefs can be written as the maximum over a set of random variables, each corresponding to the movement in the evaluator's beliefs for a given choice of covariates to reveal. The proof proceeds by first reducing this problem to studying the maximum of a growing sequence of (appropriately constructed) i.i.d. variables, and then applying a result from Chernozhukov et al. (2013) to show that this maximum concentrates on its expectation as the number of covariates grows large. We conclude by bounding this expectation and demonstrating that it vanishes.

We next apply this result to compare the agent's expected payoffs under human and algorithmic evaluation, under an assumption that the algorithm processes a sufficiently larger fraction of covariates compared to the human (in a sense we make precise). We show that for rich enough covariate spaces, agents with convex payoffs robustly prefer algorithmic evaluation, while agents with concave payoffs robustly prefer human evaluation. In the banking example, this means that a risk-averse agent whose payoffs are increasing in the bank's assessment of repayment probability should prefer the traditional bank (when the covariate space is rich). By contrast, an agent who benefits from the bank having an accurate assessment—for instance, if he only wants a loan when his business will succeed and he can repay—should prefer the fintech bank (when the covariate space is rich). We view these conclusions as relevant not only in a far limit of many covariates, and provide a bound for the number of covariates that is needed for our result to hold. In a simple binary example where the human institution observes 10% of covariates, and the algorithmic institution observes 90% of covariates, then our result holds as long as there are at least 10 covariates.

We subsequently strengthen our main results in four ways: First, we show that not only does the expected value of context vanish for each agent, but in fact the expected *maximum* value of context across agents also vanishes. Second, we show that our main results extend when the agent and evaluator interact in a disclosure game, where the agent chooses which nonstandard covariates to reveal, and the evaluator makes inferences about the agent based on which covariates are revealed (given the agent's equilibrium reporting strategy). Third, we provide an abstract learning condition under which our results extend: It is enough for the informativeness of each individual set of covariates to vanish as the total number of covariates grows large. Finally, we show that if the agent has some knowledge about which nonstandard covariates are likely to be predictive, then the limiting value of context is simply what the agent could gain by announcing his known context, prior to the realization of the prediction problem.

Our model is not meant to be a complete description of the differences between human and algorithmic evaluation. For example, we do not consider human or algorithmic bias (Kleinberg et al., 2017; Gillis et al., 2021), explainability (Yang et al., 2024), preferences for empathetic evaluators, or the possibility that the human evaluator has access to information that is not available to the algorithm (e.g., for privacy protection reasons as in Agarwal et al. (2023)). We also suppose that both evaluators form correct conditional expectations, thus abstracting away from the possibility of algorithmic overfitting and of bounded human rationality (e.g., as considered in Spiegler (2020) and Haghtalab et al. (2021)).[2] We leave extensions of our model that include these other interesting differences to future work.

1.1 Related Literature

Our paper is situated at the intersection of the literatures on learning (Section 1.1.1) and strategic information disclosure (Section 1.1.2), where our analysis is primarily differentiated from the previous frameworks by our assumption that the agent has model uncertainty (see Section 1.1.1). Our paper is also inspired by a recent empirical literature that compares human and AI evaluation, which we review in Section 1.1.3.

[2]The problem of overfitting, while practically important, is a function of how the algorithm is trained. We are interested here in intrinsic differences between the qualitative nature of human and algorithmic evaluation, which are difficult to resolve by training the algorithm differently.

1.1.1 (Asymptotic) Learning

A large literature studies asymptotic learning and agreement across Bayesian agents (Blackwell and Dubins, 1962). Our main result (Theorem 1) can be viewed as bounding (in expectation) the differences in beliefs across Bayesian agents who are given different information. As in Vives (1992), Golub and Jackson (2012), Liang and Mu (2019), Harel et al. (2020), and Frick et al. (2023) among others, we quantify the rate of convergence in beliefs. The learning rates that we look at are, however, of a different nature from those studied previously. One important distinction is that these previous papers consider asymptotics as the total amount of information accumulates, while our analysis considers asymptotics with respect to a sequence of information structures that we show are increasingly less informative. A second important difference is that the classic learning models suppose that the agent updates to a signal with a known signal structure, while our agent is uncertain about the signal structure (as in Acemoglu et al. (2015) and Morris and Yildiz (2019)). Our results characterize the informativeness of this signal in expectation, where the agent's model uncertainty takes a particular (and new) form motivated by the applications we have in mind.

Finally, our paper is related to Di Tillio et al. (2021), which compares the informativeness of an unbiased signal to the informativeness of a selected signal whose realization is the maximum realization across i.i.d. unbiased signals. Again the key difference is our assumption of model uncertainty—that is, in Di Tillio et al. (2021), the signal structures that are being compared are deterministic and known, while in ours they are random and compared in expectation. In particular, our agent's prior belief over signal structures can have support on signal processes which are not i.i.d. (for example, it may be that the meaning of one signal is dependent on the meaning of another).

1.1.2 Strategic Information Disclosure

Several literatures study persuasion via strategic information disclosure. Our model—in which the sender has private information about his covariate vector, and selectively chooses which elements to disclose to a naive receiver—is closest to models of disclosure of hard information (Dye, 1985; Grossman and Hart, 1980), in particular Milgrom (1981).[3] The key

[3] A similar model of information is considered in Glazer and Rubinstein (2004) and Antic and Chakraborty (2023).

difference (which follows from our assumption of model uncertainty) is that our sender has uncertainty about how his reports are interpreted. Additionally, our focus is not on examining which incentive-compatible reporting strategy is optimal,[4] but instead on asymptotic limits of belief manipulability as the number of components in the covariate vector grows large. This latter focus is special to our motivating applications.

Our model also has important differences from the other main strands of the persuasion literature. Unlike models of cheap talk (Crawford and Sobel, 1982), our agent chooses between messages whose meanings are fixed exogenously (through the realization of the joint distribution relating covariates to the type) rather than in an equilibrium. Unlike the literature on Bayesian persuasion (Kamenica and Gentzkow (2011)), our sender chooses which signal realization to share ex-post from a finite set of signal realizations, rather than committing to a flexibly chosen information structure ex-ante.[5] Indeed, our model gives the sender substantial power to influence the receiver's beliefs relative to this previous literature. It is perhaps surprising, then, that despite the lack of constraints imposed on the sender, we find that the sender is extremely limited in his influence. In our model, this emerges because the sender has a limited choice from a set of information structures, whose informativeness (we show) is vanishing in the total number of covariates.[6]

1.1.3 Human vs AI Evaluation

Recent empirical papers compare the accuracy of human evaluation with AI evaluation, finding that machine learning algorithms outperform experts in problems including medical diagnosis (Rajpurkar et al., 2017; Jung et al., 2017; Agarwal et al., 2023), prediction of pretrial misconduct (Kleinberg et al., 2017; Angelova et al., 2022), and prediction of worker productivity (Chalfin et al., 2016). Nonetheless, many individuals continue to distrust al-

[4]Indeed, in our main model we do not require choice of an incentive-compatible reporting strategy, since the receiver updates to the sender's disclosure as if it were exogenous information. This is primarily for convenience—we show in Section 5.2 that our results extend in a disclosure game.

[5]Thus, for example, Bayes plausibility is not satisfied in our setting—the sender's expectation of the receiver's expectation of the state (following disclosure) is generally not the prior expectation of the state.

[6]The covariates in our model play a similar role to attributes, although the literature on attributes has focused on choice of which attributes to learn about (e.g., Klabjan et al. (2014) and Liang et al. (2022)), rather than which attributes to disclose for the purpose of persuasion. An exception is Bardhi (2023), who studies a principal-agent problem in which a principal selectively samples attributes to influence an agent's decision.

gorithmic predictions (Jussupow and Heinzl, 2020; **?**). These findings motivate our goal of understanding whether individuals should prefer human evaluators, and when instead the replacement of human evaluation with algorithmic evaluation is welfare-improving for users, as suggested in Obermeyer and Emanuel (2016) among others.

In principle, human decision-making guided by algorithmic predictions should be superior to either human or algorithmic prediction alone. In practice the evidence is more mixed, with the provision of algorithmic recommendations sometimes leading human decision-makers to less accurate predictions (Hoffman et al., 2017; Angelova et al., 2022; Agarwal et al., 2023).[7] The question of how to aggregate human and machine evaluations is thus important but subtle, and depends on (among other things) whether human decision-makers understand the correlation between their information and that of the algorithm (McLaughlin and Spiess, 2022; Gillis et al., 2021; Agarwal et al., 2023). We abstract away from these complexities, focusing instead on (one aspect of) the more basic question of why human oversight is even necessary to begin with. We provide a tractable way of formalizing the advantage of human evaluation, and quantify the size of this advantage.

2 Model

2.1 Setting

Let $\mathcal{X}_n = \{0,1\}^n$ denote a set of binary covariate vectors, which are uniformly distributed in the population.[8] A *prediction problem* is a pair $(\mathcal{Y}, f)$ consisting of a bounded set of types $\mathcal{Y} \subseteq [-\overline{y}, \overline{y}]$ and a mapping $f : \mathcal{X}_n \to \mathcal{Y}$ from covariate vectors to types. To ease exposition we will simply refer to f as the prediction problem going forward.

An agent with covariate vector $x \in \mathcal{X}_n$ chooses between two institutions, one in which predictions are made by humans (the *human institution*) and another in which they are made by algorithms (the *algorithmic institution*). The agent knows his covariate vector and the distribution F over prediction problems, but not which specific prediction problem $f \sim F$ will realize.

[7] Other papers instead consider algorithmic prediction tools that take human evaluation as an input, with greater success towards improving accuracy (e.g., Raghu et al. (2019)).

[8] All of our results extend for arbitrary finite sets $\mathcal{X}_n$.

Example 1 (Bank Underwriting). A small business owner chooses between signing a multi-year relationship-lending contract with Bank H, where loan officers conduct underwriting, or FinTech A, which relies on an automated credit-scoring model. At the time of contracting, the owner cannot foresee which financing tools he will need in the future; for example, whether he will need a revolving working-capital line to finance an order surge or a multi-year equipment loan to replace critical machinery. When the firm requires financing, the evaluation is either handled by human underwriters at Bank H or FinTech A's algorithm according to the business owner's selected contract.

Example 2 (Medical Diagnosis). The agent chooses between Plan H, a traditional health insurance plan covering care at hospitals with in-person physicians, and Plan A, a new health insurance plan that covers care by digital health platforms. The enrollment choice is binding for several years, and is made before the patient knows which condition will arise, e.g., whether he will need to be evaluated for a heart arrhythmia condition or a thyroid disorder. Once the condition manifests, diagnosis is carried out according to the chosen plan.

The agent's payoffs $u(\hat{y}, y)$ depend on the prediction of his type $\hat{y}$ (described in the following section) and his true type y, as captured by a Lipschitz continuous utility function $u : [-\overline{y}, \overline{y}]^2 \to \mathbb{R}$.

2.2 Human versus Algorithmic Predictions

At both institutions, the prediction of the agent's type y is based on a set of observed covariates. Formally, for any set of covariate indices $I \subseteq \{1, \ldots, n\}$ and pair of vectors $x, x' \in \mathcal{X}_n$, define $x \sim_I x'$ if $x_i = x'_i$ for all $i \in I$, i.e., x and x' are observationally equivalent given the covariates in I. Further let $\Pi_I(x)$ denote x's equivalence class under $\sim_I$, i.e.,

$$\Pi_I(x) = \{x' \in \mathcal{X}_n \mid x' \sim_I x\} \quad \forall x \in \mathcal{X}_n.$$

Recalling that covariates are uniformly distributed in the population, the average type within x's $\sim_I$-equivalence class is given as follows.

Definition 1 (Prediction Based on I)**.** *For any covariate vector $x \in \mathcal{X}_n$ and set of covariate indices $I \subseteq \{1, \ldots, n\}$, define $\hat{f}_I(x) := \frac{1}{|\Pi_I(x)|} \sum_{x' \in \Pi_I(x)} f(x')$.*

For instance, if $f(x)$ is the agent's probability of loan repayment, then $\hat{f}_I(x)$ is the average repayment probability among individuals who are indistinguishable from the agent given their covariates in I.

Covariates are separated into two categories. *Standard* covariates, indexed by $\mathcal{S} = \{1, \ldots, s\}$, correspond to attributes with well-understood implications that are routinely collected for prediction. *Nonstandard* covariates, indexed by $\mathcal{N} = \{s+1, \ldots, n\}$, correspond to all other relevant attributes of the agent. In Example 1, standard covariates might include credit scores, outstanding debts, and income and employment status, while nonstandard covariates might include seasonal cash-flow patterns, customer reviews, point-of-sale transaction data, supplier stability, and the business owner's community reputation. In Example 2, standard covariates might include prior diagnoses, family medical history, lab tests and imaging results, while nonstandard covariates might include the patient's religious practices, genetic data, wearable device data, and financial circumstances.[9]

Both institutions collect the standard covariates S, but they differ in which nonstandard covariates they collect.[10] The algorithmic institution supplements the standard covariates with a set of nonstandard covariates $A \subseteq \mathcal{N}$. Specifically, the set A includes a fraction α_b of the nonstandard covariates, drawn from an arbitrary distribution on $\mathcal{N}$ that is independent of the agent's covariate vector x and the prediction problem f. The algorithm's prediction is then $\hat{f}_{S\cup A}(x)$ for the realized A. To simplify notation, for any set of nonstandard covariates $N \subseteq \mathcal{N}$, henceforth let

$$U_x^f(N) = u\left(\hat{f}_{S\cup N}(x), f(x)\right)$$

denote the agent's payoff when his true type is $f(x)$ and the prediction about his type is $\hat{f}_{S\cup N}(x)$. So the agent's expected payoffs at the algorithmic institution are $\mathbb{E}_A[U_x^f(A)]$.

At the human institution, the set S is instead supplemented by a personalized set $H_{x,f} \subseteq \mathcal{N}$. This set includes at most a fraction α_h of the nonstandard covariates, where $\alpha_h < \alpha_b$, reflecting that the human evaluator cannot process as many inputs as the algorithm can. We interpret $H_{x,f}$ as a small but personalized set of covariates revealed to the human evaluator during an in-person interaction. The human evaluator's prediction for the

[9] See Acosta et al. (2022) for further examples of nonstandard patient covariates that may be predictive, but which are not currently used by clinicians for medical evaluations.

[10] Our results extend without change if the institutions collect a f-dependent subset of standard covariates, $I_f \subseteq S$.

agent with covariate vector x is $\hat{f}_{S\cup H_{x,f}}(x)$, and the agent's payoffs are $U_x^f(H_{x,f})$. Figure 1 summarizes the model.

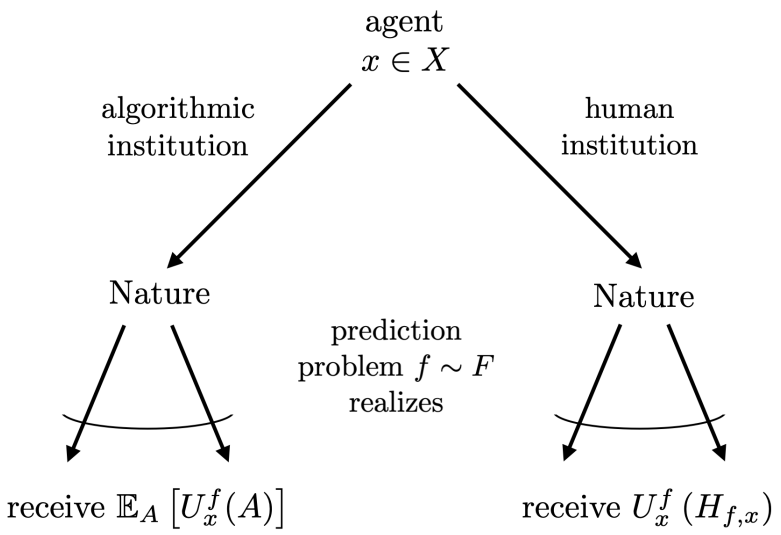

Figure 1: The agent commits to either the algorithmic or human institution before the prediction problem f is realized. Once f is known, the agent is evaluated by the chosen institution, where the algorithmic institution considers the random but standardized set of nonstandard covariates A, while the human institution considers the targeted set $H_{x,f}$.

Rather than specifying a particular form for how the set of covariates $H_{x,f}$ depends on (x, f), we consider lower and upper bounds for the agent's payoffs at the human institution. Let $\mathcal{H}_n$ denote all subsets of $\mathcal{N}$ whose size does not exceed $\alpha_h n$ (i.e., which respect the human capacity constraint).

Definition 2. *Say that the agent with covariate vector* x robustly prefers the algorithm *if*

$$\mathbb{E}_{f,A}\left[U_x^f(A)\right] > \mathbb{E}_f\left[\max_{H\in\mathcal{H}_n} U_x^f(H)\right] \tag{1}$$

and robustly prefers the human *if*

$$\mathbb{E}_{f,A}\left[U_x^f(A)\right] < \mathbb{E}_f\left[\min_{H\in\mathcal{H}_n} U_x^f(H)\right] \tag{2}$$

The first part of Definition 2 compares the agent's expected payoff under algorithmic evaluation with the *best-case* expected payoff under human evaluation, namely when the human evaluator supplements the standard covariates in S with those (up to) $\alpha_h \cdot n$ nonstandard covariates that maximize the agent's payoffs. If the agent's expected payoff is higher under

algorithmic evaluation than under any human selection rule (i.e., any map from (x, f) to a set of nonstandard covariates $H_{x,f}$), we say that the agent *robustly prefers to be evaluated by the algorithm*. The second part of the definition compares the agent's expected payoff under algorithmic evaluation with the *worst-case* expected payoff under human evaluation, namely when the human evaluator observes those (up to) $\alpha_h \cdot n$ covariates that minimize the agent's payoffs. If every human selection rule yields a higher payoff than algorithmic evaluation, then we say that the agent *robustly prefers to be evaluated by the human*.[11] In practice, we would expect the set of revealed covariates $H_{x,f}$ to lie somewhere between these extremes,[12] but if we can rank the institutions under Definition 2 then that same ranking would hold for any other model of $H_{x,f}$.

2.3 Uncertainty over the Prediction Problem f

When choosing between institutions, the agent does not know which prediction problem f will materialize, but holds a prior belief $f \sim F$ satisfying the following condition.

Assumption 1 (Symmetry)**.** *For every fixed vector of standard covariates $x_S \in \{0,1\}^s$, there is a distribution $\pi_{x_S} \in \Delta([-\overline{y}, \overline{y}])$ (independent of n) such that*

$$f(x_S, x_N) \overset{iid}{\sim} \pi_{x_S}$$

across all vectors of nonstandard covariates $x_N \in \{0,1\}^{n-s}$.[13]

The distribution π_{x_S} over outcomes can be interpreted as the conditional belief about the agent's type given his covariates, but *without* knowledge of the prediction problem. This assumption states that the agent's type has the same distribution for all possible values of the agent's nonstandard covariates (although it can depend on the standard covariates).

Example 3. Suppose $n = 2$ where x_1 is standard while x_2 is nonstandard. Then there are

[11] In Section 5.2 we further discuss the extent to which these interpretations are valid when the evaluator also updates her beliefs to the selection of covariates.

[12] Angelova et al. (2022) provide evidence that some judges condition on irrelevant defendant covariates when predicting misconduct rates.

[13] All of our results extend if this i.i.d. assumption is replaced by an assumption of exchangeability.

four possible covariate vectors.

x_1	x_2	$f(x_1, x_2)$
0	0	$f(0,0)$
0	1	$f(0,1)$
1	0	$f(1,0)$
1	1	$f(1,1)$

The assumption says that $f(0,0)$ and $f(0,1)$ are iid under the agent's prior, while $f(1,0)$ and $f(1,1)$ are iid under the agent's prior.

The content of this assumption can be described in two parts. First, because the distribution of $f(x_S, x_N)$ does not depend on x_N, the agent is agnostic about which nonstandard covariates will matter and how they will matter in the realized prediction problem. For example, before a patient knows which medical condition is relevant, he does not know whether it will be more informative for his genetic profile or travel history to be queried, or in what direction these features will influence the assessment.

Second, because the distribution π_{x_S} does not depend on n, the total number of covariates is not a measure of the amount of information. Instead, it measures the richness of the informational environment and the potential complexity of the mapping f. When n is small, the type y is completely determined from a small set of covariates, while when n is large, the type potentially depends on many covariates. Under Assumption 1, growing n expands the space of admissible mappings; that is, the type itself is no more or less predictable, but the form that the predictive rule takes becomes potentially richer and more complex.[14]

Assumption 1 is crucially placed *ex-ante* on the agent's prior, and not *ex-post* on the realized prediction problem f. For example, the function $f(x_1, \ldots, x_n) = x_1$, which says that the only covariate that matters is x_1, is strongly asymmetric (x_1 is differentiated from the other covariates) and also features a single "large" covariate (the realization of x_1 completely determines y). Our assumptions do not rule out the possibility of this function, and indeed one leading example is the one in which all prediction problems are possible and equally likely.

[14] As n grows large, the smallest possible informational size of each covariate (in the sense of McLean and Postlewaite (2002)) vanishes. But we do not require each covariate to be equally informationally relevant in the realized function. So, for example, $f(x_1, \ldots, x_n) = x_1$ can be in the support of the agent's beliefs for n arbitrarily large (see Example 4).

Example 4. Let $y \in \{0,1\}$, in which case the space of possible prediction problems $f : \mathcal{X}_n \to \mathcal{Y}$ can be identified with $\{0,1\}^{2^n}$. Suppose that for each n, the agent has a uniform prior on the set of all functions $\{0,1\}^{2^n}$. Then Assumption 1 is satisfied.

In contrast, a simple example that is ruled out is when the outcome is perfectly predictable from some small set of covariates.

Example 5. The type is equal to the value of the nonstandard covariate x_i, where the index i is drawn uniformly at random from $\mathcal{N}$.

In Section 5, we extend our main results under different relaxations of Assumption 1.

Having placed this assumption, the standard covariates will no longer play any important role in the analysis, so going forward we simplify exposition by setting $S = \varnothing$ and $\mathcal{N} = \{1, \ldots, n\}$.

2.4 Discussion of Model

The interpretation of "algorithm" and "human." In our model, a key distinction between human and algorithmic evaluation is that the human can adapt which covariates are acquired based on other properties of the agent, while the algorithm is based on a pre-specified set of covariates. This is an appropriate description of the machine learning algorithms typically deployed in the applications we have mentioned, which are usually supervised machine learning algorithms pre-trained on a large data set. But new machine learning algorithms, such as large language models, blur this distinction, and future evaluations (e.g., medical diagnoses) may be conducted by black box systems with which the agent can communicate. From this more forward-looking perspective, our results can be understood as comparing the merits of online versus offline learning. That is, is it better to have an evaluator dynamically acquire information given feedback from the agent, or to learn from a larger pre-specified set of covariates?

Strategic Disclosure. In Section 2.2, we interpret the quantities $\mathbb{E}_f\left[\min_{H \in \mathcal{H}} U_x^f(H)\right]$ and $\mathbb{E}_f\left[\max_{H \in \mathcal{H}} U_x^f(H)\right]$ as lower and upper bounds for the agent's expected payoff at the human institution. This is a suitable interpretation when covariate selection reveals no information about the agent's type—for instance, when the evaluator chooses H or updates her beliefs as

if H were chosen exogenously.[15] However, if the agent strategically chooses which covariates to reveal in a disclosure game and the evaluator updates her beliefs with respect to the agent's disclosure strategy, then whether these quantities bound the agents' payoffs depends on the agent's information at the time of disclosure.[16] In Section 5.2, we extend our model to incorporate strategic disclosure and develop alternative lower and upper bounds suited to that model. We show that our main results extend.

Human oversight of algorithms. Our stark separation between algorithmic and human institutions is deliberately stylized, and intended to clarify two contrasting modes of evaluation. In practice, many institutions adopt hybrid models, where AI provides guidance but a human oversees the algorithm. While our analysis does not focus on this case, one natural extension would compare (i) AI predictions based on a large set of standardized covariates with (ii) AI–human predictions that supplement the covariates in (i) with a small, targeted set of additional covariates chosen by the human. Taking the example from the introduction, a human loan officer could realize that the borrower's employment status was misleading in the algorithm, and condition further on the borrower's involvement in a startup. A straightforward implication of Theorem 1 is that the expected value of such targeted additions vanishes as the covariate space grows. In other words, although algorithms can occasionally be improved by incorporating new inputs, the measure of such cases is small.

3 The Value of Context

A key input towards understanding the comparison between the human and algorithmic institutions is quantifying the extent to which predictions are responsive to individualized context. This section formalizes a measure for this responsiveness and proves a result about its importance when the set of covariates is rich.

[15] Jin et al. (2021) and Farina et al. (2023) find that the beliefs of experimental subjects fall somewhere in between this naive benchmark and equilibrium beliefs, since subjects do not completely account for the strategic nature of disclosure.

[16] This is because the agent may convey information about undisclosed covariates through the choice of which covariates to reveal; for example, only revealing x_1 if all of the remaining covariates are equal to 1.

3.1 Definition and Examples

Definition 3 (Value of Context)**.** *For any $n \in \mathbb{Z}_+$, prediction problem $f : \mathcal{X}_n \to \mathcal{Y}$, and covariate vector $x \in \mathcal{X}_n$, the* value of context *is*

$$v_n(f, x) = \max_{H \in \mathcal{H}_n} U_x^f(H) - U_x^f(\varnothing)$$

i.e., the best possible improvement in the agent's utility when the evaluator additionally observes up to $\alpha_h \cdot n$ covariates for the agent.

Whether the value of context is low or high depends on the structure of the prediction problem and how it interacts with the agent's covariate vector. We construct examples below where the value of context can be as large as $\overline{y}$ (the upper bound on payoffs) and when it is arbitrarily small.

Example 6 (The Value of Context is High). Let $u(\hat{y}, y) = \hat{y}$, i.e., the agent's payoff is the predicted type. There are $n > 3$ covariates $x_1, \ldots, x_n$, where x_1 is standard and $x_2, \ldots, x_n$ are nonstandard. The prediction problem is

$$f(x_1, \ldots, x_n) = \begin{cases} \overline{y} & \text{if } x_1 = x_2 \\ -\overline{y} & \text{if } x_1 \neq x_2 \end{cases}$$

That is, the meaning of the standard covariate is determined by the realization of a particular nonstandard covariate.

Both institutions observe x_1. The human institution additionally observes up to one covariate, while the algorithmic institution observes 90 covariates chosen uniformly at random from $\{x_2, \ldots, x_n\}$. The agent's covariate vector is the vector of all ones.

Then the agent's payoff is $U_x^f(\varnothing) = 0$ when the prediction is conditioned only on the standard covariate x_1, and it is $U_x^f(\{2\}) = \overline{y}$ when the evaluator additionally observes the best additional nonstandard covariate, x_2. So the value of context is $\overline{y}$. This example illustrates settings where the correct interpretation of routinely collected covariates hinges on a single contextual variable, so that the targeted acquisition of relevant context substantially improves the agent's payoffs.

Example 7 (The Value of Context is Low). Consider the above example with the alternative prediction problem

$$f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Then the agent's payoff is $U_x^f(\varnothing) = \left(\frac{1}{n}\right)\cdot 1 + \left(\frac{n-1}{n}\right)\left(\frac{1}{2}\right)$ when the prediction is conditioned only on the standard covariate x_1, and it is $\left(\frac{2}{n}\right)\cdot 1 + \left(\frac{n-2}{n}\right)\left(\frac{1}{2}\right)$ if the prediction is additionally conditioned on any nonstandard covariate. So the value of context is $\frac{1}{2n}$. This example illustrates settings in which the outcome aggregates many equal contributions from different covariates, so that the value of targeted acquisition is small.

Taking an expectation over the random prediction problem $f \sim F$ that the agent will face, we obtain the following ex-ante version of the value of context.

Definition 4 (Expected Value of Context)**.** *For every $n \in \mathbb{Z}_+$ and $x \in \mathcal{X}_n$, the* expected value of context *is* $V(n,x) \equiv \mathbb{E}\left[v_n(f,x)\right]$.

This quantity tells us the extent to which context can improve the agent's payoffs in expectation.

3.2 The Expected Value of Context Vanishes

Our main result says that as the covariate space becomes increasingly rich (i.e., as n grows large), the expected value of context converges to zero for every agent.

Theorem 1. *Suppose Assumption 1 holds. Then for every $\varepsilon > 0$ there is an N_ε such that*

$$V(n,x) < \varepsilon \quad \forall x \in \mathcal{X}_n, n > N_\varepsilon$$

Thus, for any fixed $\varepsilon > 0$, there exists a covariate dimension N_ε such that once the number of potential covariates exceeds this threshold ($n > N_\varepsilon$), then the expected benefit that any agent x obtains from the possibility to provide context—as measured by $V(n,x)$—is no larger than ε.

This result says that although the value of context may be large for certain realizations of the prediction problems (such as Example 6), it does not matter in expectation. Intuitively, as n grows large, prediction problems in which a small set of covariates are decisive become increasingly rare under Assumption 1, so that—with bounded payoff functions—the value of context is small *on average* across these problems. This also implies that for sufficiently large n, the provision of context does not "typically" matter; that is, the probability that the agent gains substantially from targeted information acquisition is small.

The analysis (discussed in further detail in Section 3.3) hinges on a comparison of two opposing forces. First, as n grows large, the number of distinct sets in $\mathcal{H}_n$ increases, expanding the set over which $\max_{H \in \mathcal{H}_n} U_x^f(H)$ is maximized. This increases the value of context. On the other hand, we show that the difference between any $U_x^f(H)$ and $U_x^f(H')$ decreases as n grows large, since the predictions $\hat{f}_H(x)$ and $\hat{f}_{H'}(x)$ are sample averages that concentrate around their common expectation. What we have to determine is whether the growth rate in the number of subsets of nonstandard covariates (of size $\leq \alpha_h n$) is sufficiently large to maintain a non-vanishing utility gap across these sets. The answer turns out to be no.

As this intuition suggests, Theorem 1 extends beyond the i.i.d. baseline in Assumption 1, since what matters is that the random prediction $\hat{f}_H(x)$ concentrates on its expectation sufficiently quickly as n grows large. We provide an abstract learning condition in Section 5.3 that formalizes this requirement. Additionally, we show in Section 5.4 that if the agent has some knowledge about the predictiveness of nonstandard covariates, then the limiting value of context is not zero but rather a quantity that reflects what the agent already knows.

3.3 Proof Sketch

The core of the proof of Theorem 1 is an argument that the extent to which context changes the evaluator's prediction in expectation vanishes in n. We outline that argument here. For each n, there are $K_n = \sum_{j=0}^{\lfloor \alpha_h n \rfloor} \binom{n}{j}$ sets of $\alpha_h n$ (or fewer) nonstandard covariates that can be disclosed. Enumerate these sets as $H_1, \ldots, H_{K_n}$. Each set H_k induces a prediction

$$Z_k^n \equiv \hat{f}_{H_k}(x) = \frac{1}{|\Pi_{H_k}(x)|} \sum_{x' \in \Pi_{H_k}(x)} Y_{x'}$$

where $Y_{x'} := f(x')$ denotes the (random) type for an agent with covariate vector x' under the (random) function f. Let $Z_\varnothing^n = \hat{f}_\varnothing(x)$ denote the prediction of the agent's type based on his standard covariates only. We show that

$$\mathbb{E}\left[\max_{1 \leq k \leq K_n} Z_k^n\right] - \mathbb{E}[Z_\varnothing^n] \to 0$$

so that context cannot (in expectation) substantially increase the prediction. After normalizing $E[Z_\varnothing^n] = 0$, this reduces to showing $\mathbb{E}[\max_{1 \leq k \leq K_n} Z_k^n] \to 0$.

There are two challenges to analyzing this quantity. First, the correlation structure of $Z_1^n, \ldots, Z_{K_n}^n$ can be complex: The variables Z_k^n are neither independent (because the

same type value Y_x can appear as an element in different sample averages $Z_k^n, Z_{k'}^n$) nor identically distributed (because the sample averages are of different sizes depending on how many nonstandard covariates are revealed). The second challenge is that the length of the sequence $(Z_1^n, \ldots, Z_{K_n}^n)$ grows exponentially in n. Thus even though each term within the maximum eventually converges to a normally distributed random variable (with shrinking variance), the errors of each term may in principle accumulate in a way that the maximum grows large.

Our approach is to first construct new i.i.d. variables $\widetilde{Z}_k^n$, with the property that

$$\mathbb{E}\left[\max\{Z_1^n, \ldots, Z_{K_n}^n\}\right] \leq \mathbb{E}\left[\max\{\widetilde{Z}_1^n, \ldots, \widetilde{Z}_{K_n}^n\}\right] \tag{3}$$

Applying a result from Chernozhukov et al. (2013), we show that $\max_{1\leq k\leq K_n} \widetilde{Z}_k^n$ (properly normalized) converges to $\max_{1\leq k\leq K_n} Z_k^{Normal}$ in distribution, where (due to properties of our problem) $Z_k^{Normal} \sim_{iid} N\left(0, \frac{1}{2^{n(1-\alpha_h)}}\right)$. Having reduced the analysis to studying the expected maximum of i.i.d. Gaussian variables, classic bounds apply to show that this quantity is bounded above by

$$\frac{1}{\sqrt{2^{n(1-\alpha_h)}}}\sqrt{\log(K_n)}. \tag{4}$$

This display quantifies the importance of each of the two forces discussed in the previous section. First, as n grows larger, the number of feasible predictions $K_n = \sum_{j=0}^{\lfloor \alpha_h n \rfloor} \binom{n}{j} \leq 2^n$ grows exponentially in n, increasing the expected value of context. But second, as n grows larger, each Z_k concentrates on its expectation, where its variance, $\frac{1}{2^{n(1-\alpha_h)}}$, decreases exponentially in n. What the bound in display (4) tells us is that the exponential growth in the number of variables is eventually dominated by the exponential reduction in the variance of each variable, yielding the result.

4 Human versus Algorithmic Institution

We now turn to the question of which institution the agent prefers.

Assumption 2. *The agent's expected utility can be written as $\mathbb{E}[\phi(\hat{y})]$ for some twice continuously differentiable function ϕ.*[17]

[17]Restricting to utility functions that depend on a posterior mean is a common assumption in the literature on information design, see e.g., Kamenica and Gentzkow (2011), Frankel (2014) and Dworczak and Martini (2019).

Theorem 2. *Suppose Assumptions 1 and 2 hold and suppose* $\alpha_b > \frac{1+\alpha_h}{2}$. *Then there exists* $N > 0$ *such that for all* $n \geq N$*:*

(a) If ϕ *is strictly convex, the agent robustly prefers the algorithmic institution;*

(b) If ϕ *is strictly concave, the agent robustly prefers the human institution.*

The case of convex ϕ (Part (a)) corresponds to a preference for more accurate evaluations.[18] Such an agent prefers for the evaluation to be based on more information (advantaging the algorithmic institution), but also prefers for the evaluation to be based on more relevant covariates (advantaging the human institution). We show that what eventually dominates is how many covariates the evaluators observe, not how they are selected; thus when the algorithmic prediction is based on sufficiently more covariates than the human prediction, an agent who prefers accuracy eventually prefers the algorithmic institution.

Part (b) of Theorem 2 says that if instead ϕ is concave, and the human observes sufficiently few covariates, then the agent eventually robustly prefers the human institution. Loosely speaking, when the human prediction is based on only a few of many covariates, then even the human's worst-case selection bears little impact on the evaluation. Examples 8 and 9 illustrate decision problems that meet the conditions of each part of the theorem.

Example 8 (Accuracy). Suppose the agent's type is $y \in \{0, 1\}$, and the evaluator chooses an action $a \in [0, 1]$ based on the observed covariates. The evaluator and agent share the utility function $-\mathbb{E}[(a - y)^2]$; that is, both would like for the action to be as accurate as possible. The evaluator's optimal action is $a = \hat{y}$, and the agent's expected payoff given this action is $\mathbb{E}\left[-(\hat{y} - y)^2\right] = \mathbb{E}\left[-\left(y(1 - \hat{y})^2 + (1 - y)(\hat{y})^2\right)\right] = \mathbb{E}\left[\phi(\hat{y})\right]$, where $\phi(\hat{y}) = \hat{y}^2 - \hat{y}$ is convex. Then Part (a) of Theorem 2 implies that the agent prefers evaluation by the algorithmic institution in rich covariate spaces.[19]

[18] Consider any two sets of covariates $A \subset A'$ and let $\hat{y}_A$, $\hat{y}_{A'}$ be the corresponding posterior expectations. The distribution of $\hat{y}_{A'}$ (i.e., the posterior expectation that conditions on more information) is a mean-preserving spread of the distribution of $\hat{y}_A$. When ϕ is convex, the former leads to a higher expected utility. Such an agent "prefers more accurate evaluations" in the sense that giving the evaluator better information (in the standard Blackwell sense) leads to an improvement in the agent's expected utility.

[19] Although the conditions of Theorem 2 are no longer met when y is not binary, we show in Appendix P.1 that the conclusion of Part (a) of Theorem 2 generalizes for arbitrary y given the mean-squared error payoff function described in this example.

Example 9 (Risk Aversion). Suppose that a borrower benefits from higher predictions of his loan repayment probability, but is risk averse over this prediction. Specifically let his utility function be $u(\hat{y}, y) = \phi(\hat{y})$ for some increasing, concave, and twice continuously-differentiable ϕ. Then Part (b) of Theorem 2 says that the agent prefers to be evaluated by the human institution in rich covariate spaces.

Theorem 2 relies on Theorem 1, and in particular on the asymptotic rate of convergence demonstrated in its proof. We briefly explain the case in which ϕ is strictly convex. Define

$$V_A(n, x) = \mathbb{E}_{f,A}[U_x^f(A) - U_x^f(\varnothing)]$$

to be an algorithmic analogue to the value of context, i.e., the expected payoff gain from algorithmic evaluation relative to revealing only the standard covariates. Then Part (a) of Definition 2 can be restated as saying that the agent robustly prefers the algorithmic evaluator if

$$V(n, x) < V_A(n, x). \tag{5}$$

As n grows large, both quantities converge to zero: $V(n, x) \to 0$ by Theorem 1 and $V_A(n, x) \to 0$ by the Law of Large Numbers. Thus the question is which converges faster. From the proof of Theorem 1, we can asymptotically upper-bound the value of context as $V(n, x) < \frac{a_n}{\sqrt{2^{(1-\alpha_h)n}}}$, where a_n grows sub-exponentially. Using the convexity of ϕ, we further establish the asymptotic lower-bound $V_A(n, x) > \frac{b}{2^{(1-\alpha_b)n}}$ for some constant b. When $\alpha_b > \frac{1+\alpha_h}{2}$, the denominator in the lower bound decreases more slowly than in the upper bound, implying that $V(n, x)$ converges to zero faster than $V_A(n, x)$.[20]

In our setting, the absolute value of evaluation—whether algorithmic or human—vanishes as the covariate space grows. This is not, however, central to our results: what matters is the relative performance of the two approaches. To illustrate, suppose instead that outcomes take the form $y = f(x) + \varepsilon_n$, where the noise term ε_n becomes increasingly concentrated as n grows. If ε_n concentrates sufficiently quickly, then both $V(n, x)$ and $V_A(n, x)$ converge to strictly positive limits rather than vanishing, and yet the comparative advantage of the algorithm persists.

[20] In the proof we go further by showing that when $\alpha_b < \frac{1+\alpha_h}{2}$ and $\phi'(\mathbb{E}[y]) \neq 0$, the implication of Theorem 2 fails to hold. The requirement $\phi'(\mathbb{E}[y]) \neq 0$ is crucial here. If $\phi'(\mathbb{E}[y]) = 0$, the value of context converges to 0 at a faster rate that what we demonstrate in Theorem 2. This implies that the agent can robustly prefer the algorithmic institution even when $\alpha_b < \frac{1+\alpha_h}{2}$.

The precise value of threshold N in Theorem 2 depends on two factors—how fast the ex-ante random evaluations $\hat{f}$ converge to their Gaussian approximations, and how fast these Gaussian counterparts converge to their mean. We separate these two factors in the result below.

Corollary 1. *Define*

$$N_f = \min\left\{ n_0 \in \mathbb{R}_+ : \mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k^n|\right] \leq \frac{\sqrt{2}}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)} \quad \forall n \geq n_0 \right\}$$

$$N_\phi = \min\left\{ n \in \mathbb{R}_+ : \left(\alpha_b - \frac{\alpha_h+1}{2}\right) n - \frac{1}{2}\log_2(n+1) > \log_2\left(2\sqrt{2}\frac{M_1}{M_2}\right) \right\}.$$

where Z_k^n is as defined in Section 3.3 and $M_1 = \max_{\hat{y}\in[-\overline{y},\overline{y}]} |\phi'(\hat{y})|$ and $M_2 = \min_{\hat{y}\in[-\overline{y},\overline{y}]} |\phi''(\hat{y})|$. The conclusions of Theorem 2 hold for any $n \geq \max\{N_f, N_\phi\}$.

The corollary identifies a finite threshold n beyond which Theorem 2 applies. The threshold is the larger of two values, N_f and N_ϕ, where N_f depends only on the agent's uncertainty over the prediction problem, while N_ϕ, depends only on the agent's utility function.

In more detail, the quantity N_f is the smallest n for which $\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k^n|\right]$ is sufficiently small for our asymptotic bound on its size (used in the proof of Theorem 2) to apply. Since we demonstrated in the proof of Theorem 1 that $\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k^n|\right]$ vanishes as n grows large, N_f is finite.

The quantity N_ϕ is determined by two statistics of the agent's utility function: M_1, the maximum slope of ϕ, and M_2, the minimum curvature (in absolute value). Intuitively, M_1 measures how sensitive ϕ is to the prediction $\hat{y}$, while M_2 measures how strongly ϕ penalizes variation. A smaller ratio M_1/M_2 implies that N_ϕ is smaller. This ratio plays a similar role to the coefficient of absolute risk aversion, which also compares the slope of the agent's payoff to its curvature.[21]

We next show that $\max\{N_f, N_\phi\}$ can be small in practice.

Example 10. As in Example 8, suppose the agent's utility is $\phi(\hat{y}) = \hat{y}^2 - \hat{y}$, and let the outcome distribution π_{x_S} be Bernoulli on $\{0, 1\}$ with mean 1/3. Using 1000 Monte Carlo draws, we estimate $\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k^n|\right]$ for each n and compare it with the asymptotic bound. This yields an estimate of $\hat{N}_f = 10$.

[21] Recall that the coefficient of absolute risk aversion of the function ϕ is $-\frac{\phi'(\hat{y})}{\phi''(\hat{y})}$.

For the utility function, $M_1 = 1$ and $M_2 = 2$, so $N_\phi = 6$. Figure 2 fixes $\alpha_h = 0.1$ and plots N_ϕ against varying values of α_b. Taken together (and assuming N_f is our estimate $\hat{N}_f$), these computations imply that if the human institution observes 10% of covariates and the algorithmic institution observes 90%, then the comparisons in Theorem 2 already hold for all $n \geq 10$.

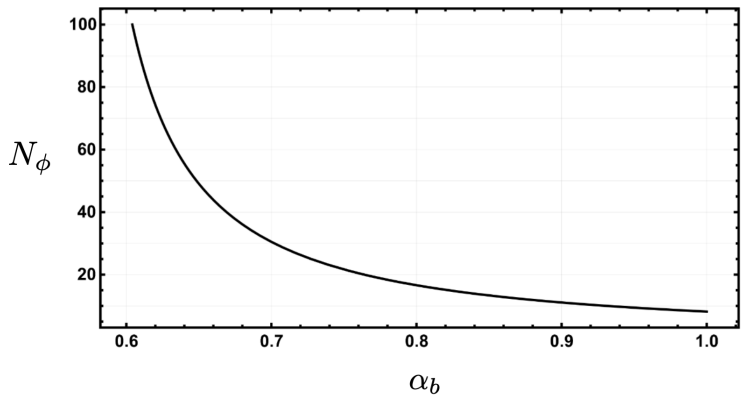

Figure 2: Let $\phi(\hat{y}) = \hat{y}^2 - \hat{y}$ and $\alpha_h = 0.1$. Then N_ϕ is depicted here as a function of α_b.

5 Extensions

We now show that we are able to strengthen our main results in the following ways. Section 5.1 shows that not only does the expected value of context vanish for each individual agent, but also the expected maximum value of context across agents vanishes. In Section 5.2, we show that our main results extend when the agent and evaluator interact in a disclosure game, wherein the evaluator updates his beliefs to the agent's strategic choice of what to disclose. Section 5.3 provides an abstract condition on the learning environment under which our main results hold, which requires the evaluator's uncertainty about the agent's type to grow sufficiently fast in n. Finally, in Section 5.4 we allow the agent to have prior knowledge about the role of certain nonstandard covariates.

5.1 Max Value of Context Across Agents

So far we have studied the expected value of context for a single agent. We next show that our results extend to the expected maximum value of context in the population, as defined below.

Definition 5. *For any $n \in \mathbb{Z}_+$, the* expected maximum value of context *is*

$$V^{max}(n) = \mathbb{E}\left[\max_{x \in \mathcal{X}_n} v_n(f, x)\right].$$

Corollary 2. *Suppose Assumption 1 holds. Then the expected maximum value of context vanishes as n grows large, i.e., $\lim_{n\to\infty} V^{max}(n) = 0$.*

Thus, the expected value of context vanishes uniformly across agents in the population. It is immediate that Theorem 1 extends in any generalization of our model in which the agent has uncertainty not only over the prediction problem f but also over his covariate vector x. In fact, the proof demonstrates that the rate of convergence is sufficiently fast that Theorem 2 extends as well.

5.2 Strategic Disclosure

So far we've remained agnostic as to whether the agent or human evaluator chooses which nonstandard covariates are observed, assuming that in either case the evaluator updates as if the covariates were revealed exogenously. We now consider a more traditional disclosure game, in which the agent chooses nonstandard covariates to disclose, and the human evaluator updates her beliefs about the agent's type given the agent's disclosure rule.

For any fixed prediction problem f, call the following an *f-context disclosure game*: There are two players, the agent and the evaluator. The prediction problem f has been realized and is common knowledge across the players.[22] The set of possible *disclosures* $\mathcal{D}$ is the set of all pairs $(H, (x_i)_{i\in H})$ consisting of a set of nonstandard covariates $H \in \mathcal{H}_n$ and values for those covariates. A disclosure $d = (H, (x'_i)_{i\in H})$ is *feasible* for an agent with covariate vector $(x_1, \ldots, x_n)$ if the disclosed covariate values are truthful, i.e., $x_i = x'_i$ for every $i \in H$.

[22] We do not interpret this assumption literally. At the other extreme where f is unknown to the agent, there is no informational content in which covariates the agent chooses to reveal, and our original interpretation applies.

The agent chooses a *disclosure strategy*, which is a map $\sigma : \mathcal{X}_n \to \mathcal{D}$ from covariate vectors to feasible disclosures. The agent then privately observes his covariate vector x and discloses $\sigma(x)$. The evaluator observes this disclosure and chooses an action $\hat{y}$. That is, the evaluator's strategy is a function $\sigma_E : \mathcal{D} \to [-\overline{y}, \overline{y}]$. The evaluator's payoff is $-(\hat{y} - y)^2$ and the agent's payoff is some function $u(\hat{y})$.

In this section we focus on pure strategy Perfect Bayesian Nash equilibria (PBE) of this game, henceforth simply *equilibria*. (A similar result holds for mixed strategy equilibria, which is demonstrated in the appendix.)

Definition 6. *Let $v^D(f, x)$ denote the highest payoff gain that an agent with covariate vector x receives in any pure-strategy equilibrium of the f-context disclosure game. The* expected maximum value of context disclosure *is*

$$V^D(n) = \mathbb{E}\left[\max_{x \in \mathcal{X}_n} v^D(f, x)\right].$$

We show that the best payoff gain that an agent can receive in any pure strategy f-context equilibrium is bounded above by the maximum value of context across agents.

Proposition 1. *Suppose Assumption 1 holds. Then for all n, $V^D(n) \leq V^{max}(n)$.*

Thus, applying Proposition 1 and Corollary 2, our previous results extend.

5.3 Sufficient Residual Uncertainty

Here we provide an abstract condition on the evaluator's learning environment, under which Theorem 1 extends.

For each n, let $\mathcal{D}_n$ denote the set of all disclosures respecting the human evaluator's capacity constraint, i.e., all pairs $(H, (x_i)_{i \in H})$ consisting of a set H with $\lfloor \alpha_h \cdot n \rfloor$ or fewer nonstandard covariates, and values $(x_i)_{i \in H}$ for those covariates. Further define $\mathcal{D} = \cup_{n \geq 1} \mathcal{D}_n$ to be the set of all disclosures.

Similarly, for each n let $\mathcal{F}_n$ be the set of all prediction problems $f : \mathcal{X}_n \to [-\overline{y}, \overline{y}]$, and define $\mathcal{F} = \cup_{n \geq 1} \mathcal{F}_n$. An *evaluation rule* is any family $\rho = (\rho_f)_{f \in \mathcal{F}}$ where each $\rho_f : \mathcal{D} \to [-\overline{y}, \overline{y}]$ maps disclosures into evaluations for the given function f. Finally, fixing any update rule ρ, number of covariates n, and disclosure $d \in \mathcal{D}_n$, let $Z_d^n = \rho_f(d)$ be the random evaluation when f is drawn from $\mathcal{F}_n$ according to the agent's prior.

We impose two assumptions below on the evaluation rule. The first says that the expected evaluation Z_d^n is equal to the prior expected type $\mu \equiv \mathbb{E}[Y]$; the second says that the distribution of the evaluation concentrates on μ sufficiently fast as the number of hidden covariates n grows large. Intuitively, the assumption requires that as the number of residual unknowns—i.e., the covariates which are predictive of the type, but are not revealed to the evaluator—grows large, the informativeness of any fixed disclosure becomes small.[23]

Assumption 3 (Unbiased)**.** $\mathbb{E}[Z_d^n] = \mu$ *for every disclosure* d*.*

Assumption 4 (Fast Concentration)**.** *For any sequence of feasible disclosures* $(d_n)_{n\geq 1}$*,*

$$Var(Z_{d_n}^n) = o\left(\frac{1}{K_n}\right)$$

where $K_n = \sum_{j=0}^{\lfloor \alpha_h n \rfloor} \binom{n}{j}$ *is the number of unique sets* $I \subseteq \{1, \ldots, n\}$ *with* $\alpha_h n$ *or fewer elements.*

These assumptions do not in general represent a weakening of our main model. Previously we studied the evaluation rule $\hat{f}$ mapping each disclosure into the conditional expectation of the agent's type, and imposed Assumption 1 on the agent's prior about f. In this model, the evaluation Z_d^n for any disclosure $d = (I, (x_i)_{i\in I})$ could be represented as a sample average consisting of $2^{n-|I|}$ elements. Assumption 3 is clearly satisfied (because the update rule is Bayesian), but one can select a sequence of disclosures (d_n) such that $Var(Z_{d_n}^n) = \frac{1}{2^{(1-\alpha_h)n}}$ (see the proof of Theorem 1 for details). Thus the speed of convergence demanded in Assumption 4 need not be met when α_h is sufficiently large.

Nevertheless, Assumption 4 identifies the qualitative property of our main setting that gave us Theorem 1: residual uncertainty must have the power to overwhelm any information revealed through disclosure. Under these assumptions, our main result extends.

Proposition 2. *Suppose Assumptions 3 and 4 hold. Then for every* $\varepsilon > 0$ *there is an* N_ε *such that* $V(n, x) < \varepsilon$ *for all* $x \in \mathcal{X}_n$ *and* $n > N_\varepsilon$*.*

Thus neither the precise symmetry imposed by Assumption 1, nor the assumption of Bayesian updating in our main model, are crucial for Theorem 1.

[23] In the limit with an uninformative disclosure, the distribution of the evaluation is degenerate at the prior expectation μ for any Bayesian updating rule.

5.4 Systematic Effects for Nonstandard Covariates

Assumption 1 imposes symmetry across all nonstandard covariates. We now relax this assumption to allow agents to possess systematic knowledge about some nonstandard covariates.

To accommodate this generalization, we reinterpret our previous notation so that $\mathcal{S}$ indexes covariates with known effects and $\mathcal{N}$ indexes covariates with unknown effects. We then consider prediction based on subsets of $\mathcal{S} \cup \mathcal{N}$, where $n = |\mathcal{S}| + |\mathcal{N}|$ denotes the total number of covariates. In a slight abuse of notation, subsequently let $U_x^f(I) = u\left(\hat{f}_I(x), f(x)\right)$. This differs from our previous notation in that I is now inclusive of all of the observed covariates.[24]

Define

$$U_n^{\max} = \mathbb{E}_f\left[\max_{x \in \mathcal{X}_n} \max_{H \subseteq S \cup \mathcal{N}, |H| \leq \alpha_h n} U_x^f(H)\right]$$

to be the expectation of the highest payoff achievable by any agent in the population when at most αn covariates are revealed to the evaluator. This choice is over both standard covariates (with known effects) and nonstandard covariates (with unknown effects), thus capturing the possibility that the agent may have prior information about the covariates are revealed after the realization of f. The following result establishes the limiting value of this quantity as n, the richness of the covariate space, grows large.

Proposition 3. *Suppose Assumptions 1 and 2 hold. Then*

$$\lim_{n \to \infty} U_n^{max} = \lim_{n \to \infty} \max_{x_S \in \{0,1\}^s} \phi\left(\mathbb{E}\left(\hat{f}_\varnothing(x_S)\right)\right)$$

where $\mathbb{E}[\hat{f}_\varnothing(x_S)]$ is the expectated prediction for an agent with standard covariates x_S, when only those standard covariates are observed.[25]

In other words, the limiting expected value of context is simply what the most favorably positioned agent could achieve by revealing their standard covariates. Thus an agent knows *a priori* that they have favorable context to share may prefer human evaluation, since this gives them the chance to share that context. But they should not expect to be able to gain

[24] Previously we had $U_x^f(I) = u\left(\hat{f}_{S \cup I}(x), f(x)\right)$.

[25] We abuse notation here slightly. For any fixed standard covariate vector x_S, $\phi\left(\mathbb{E}\left[\hat{f}_\varnothing(x_S, x_N)\right]\right) = \phi\left(\mathbb{E}\left[\hat{f}_\varnothing(x_S, x'_N)\right]\right)$ for any two $x_N, x'_N \in \mathcal{N}$, justifying our simplified notation $\hat{f}_\varnothing(x_S)$.

additional advantage *ex post* by revealing context tailored to the realized prediction problem. This finding reinforces our central message that *unexpected* context—information whose relevance could not be anticipated ex ante—has diminishing value in complex informational environments. While agents may benefit from revealing favorable context they know to be relevant, the capacity to leverage unanticipated context becomes increasingly limited as the complexity of the environment grows large.

6 Conclusion

One argument against replacing human experts with algorithms is that no matter how large the set of algorithmic inputs, the set of potentially relevant circumstances and characteristics is still more numerous. In cases where some important fact is missed by a human evaluator, it is often possible to correct this oversight. There is no such safety net with an algorithmic institution.

This is a compelling narrative, yet our results suggest that it may be less important than it initially seems. When there is a large number of nonstandard covariates that may matter for the prediction problem, but the agent does not know which specific prediction problem will be relevant, then the expected value of disclosing additional information is small—even when we assume that the agent can identify the most useful covariates to disclose, and that the claims about these covariates are taken at face value.

In contrast, if the agent has substantial prior knowledge about the predictive roles of the nonstandard covariates, then our conclusion will not be appropriate. In particular, if there is a small set of covariates that predict the type and can be fully disclosed (as in Example 5), or if some nonstandard covariates have known effects (as in Section 5.4), then the expected value of disclosing additional information may be large. We thus view our results as revealing that the value of targeted information acquisition depends critically on the extent of available "structural information" about the numerous covariates that might serve as explanations.

References

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2015): "Fragility of Asymptotic Agreement under Bayesian Learning," *Theoretical Economics*, 11, 187–225.

Acosta, J., G. Falcone, P. Rajpurkar, and E. Topol (2022): "Multimodal biomedical AI," *Nature Medicine*, 28, 1773–1784.

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023): "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working Paper 31422, National Bureau of Economic Research.

Akbarpour, M., S. Malladi, and A. Saberi (2024): "Just a Few Seeds More: Value of Network Information for Diffusion," Working Paper.

Angelova, V., W. Dobbie, , and C. S. Yang (2022): "Algorithmic Recommendations and Human Discretion," Working Paper.

Antic, N. and A. Chakraborty (2023): "Selected Facts," Working Paper.

Arnold, B. C. and R. A. Groeneveld (1979): "Bounds on expectations of linear systematic statistics based on dependent samples," *The Annals of Statistics*, 220–223.

Bardhi, A. (2023): "Attributes: Selective Learning and Influence," Working Paper.

Berman, S. M. (1964): "Limit Theorems for the Maximum Term in Stationary Sequences," *The Annals of Mathematical Statistics*, 35, 502 – 516.

Blackwell, D. and L. Dubins (1962): "Merging of Opinions with Increasing Information," *The Annals of Mathematical Statistics*.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan (2016): "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, 106, 124–27.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," .

Crawford, V. P. and J. Sobel (1982): "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1431–1451.

Di Tillio, A., M. Ottaviani, and P. N. Sørensen (2021): "Strategic Sample Selection," *Econometrica*, 89, 911–953.

Dworczak, P. and G. Martini (2019): "The Simple Economics of Optimal Persuasion," *Journal of Political Economy*, 127, 1993–2048.

Dye, R. A. (1985): "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, 23, 123–145.

Farina, A., G. Frechette, A. Lizzeri, and J. Perego (2023): "The Selective Disclosure of Evidence: An Experiment," Working Paper.

Frankel, A. (2014): "Aligned Delegation," *American Economic Review*, 104, 66–83.

Frick, M., R. Iijima, and Y. Ishii (2023): "Learning Efficiency of Multiagent Information Structures," *Journal of Political Economy*, 131, 3377–3414.

Gillis, T., B. McLaughlin, and J. Spiess (2021): "On the Fairness of Machine-Assisted Human Decisions," Working Paper.

Glazer, J. and A. Rubinstein (2004): "On optimal rules of persuasion," *Econometrica*, 72, 1715–1736.

Golub, B. and M. Jackson (2012): "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *The Quarterly Journal of Economics*, 127, 1287–1338.

Grossman, S. J. and O. D. Hart (1980): "Disclosure Laws and Takeover Bids," *The Journal of Finance*, 35, 323–334.

Haghtalab, N., M. Jackson, and A. Procaccia (2021): "Belief polarization in a complex world: A learning theory perspective," *PNAS*, 118, 141–73.

Harel, M., E. Mossel, P. Strack, and O. Tamuz (2020): "Rational Groupthink*," *The Quarterly Journal of Economics*, 136, 621–668.

Hoffman, M., L. B. Kahn, and D. Li (2017): "Discretion in Hiring*," *The Quarterly Journal of Economics*, 133, 765–800.

Jin, G. Z., M. Luca, and D. Martin (2021): "Is No News (Perceived As) Bad News? An Experimental Investigation of Information Disclosure," *American Economic Journal: Microeconomics*, 13, 141–73.

Jung, J., C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein (2017): "Simple rules for complex decisions," Working Paper.

Jussupow, Ekaterina; Benbasat, I. and A. Heinzl (2020): "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion," in *In Proceedings of the 28th European Conference on Information Systems*.

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

Klabjan, D., W. Olszewski, and A. Wolinsky (2014): "Attributes," *Games and*

Economic Behavior, 88, 190–206.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133, 237–293.

Liang, A. and X. Mu (2019): "Complementary Information and Learning Traps*," *The Quarterly Journal of Economics*, 135, 389–448.

Liang, A., X. Mu, and V. Syrgkanis (2022): "Dynamically Aggregating Diverse Information," *Econometrica*, 90, 47–80.

Longoni, C., A. Bonezzi, and C. K. Morewedge (2019): "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research*, 46, 629–650.

McLaughlin, B. and J. Spiess (2022): "Algorithmic Assistance with Recommendation-Dependent Preferences," Working Paper.

McLean, R. and A. Postlewaite (2002): "Informational Size and Incentive Compatibility," *Econometrica*, 70, 2421–2453.

Milgrom, P. R. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12, 380–391.

Morris, S. and M. Yildiz (2019): "Crises: Equilibrium Shifts and Large Shocks," *American Economic Review*, 109, 2823–54.

Obermeyer, Z. and E. J. Emanuel (2016): "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine," *The New England Journal of Medicine*, 375, 1216–9.

Raghu, M., K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan (2019): "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort," Working Paper.

Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng (2017): "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Working Paper.

Spiegler, R. (2020): "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics*, 12, 81–106.

Vives, X. (1992): "How Fast do Rational Agents Learn?" *Review of Economic Studies*, 60, 329–347.

Yang, K. H., N. Yoder, and A. Zentefis (2024): "Explaining Models," Working Paper.

A Proof of Theorem 1

To simplify the exposition of the proof, we subsequently set $|S| = s$ and $|\mathcal{N}| = n$, i.e., there are s standard covariates and n nonstandard covariates.

A.1 Preliminaries and Notation

Fix any $n \in \mathbb{Z}_+$ and any covariate vector $x_n \in \{0,1\}^n$. After observing the agent's standard covariates $x_S = (x_1, \ldots, x_s)$, the evaluator's belief about the agent's full covariate vector has support on the 2^n covariate vectors whose first s entries are x_S. Let these covariate vectors be labeled by x^j where $j = 1, \ldots, 2^n$, and define $Y_j = f(x^j)$ to be the (random) type given covariate vector x^j. By Assumption 1, there is some distribution $\pi_{x_S} \in \Delta([-\overline{y}, \overline{y}])$ such that $Y_j \sim_{iid} \pi_{x_S}$.

Recall from the main text that $\mathcal{H}_n$ denotes the set of all feasible disclosures, i.e., all subsets of nonstandard covariates whose size does not exceed $h_n = \lfloor \alpha_h n \rfloor$. There are $K_n = \sum_{k=0}^{h_n} \binom{n}{k}$ such subsets, which we can enumerate as $H_1, \ldots, H_{K_n}$. For each H_k, let $S_k = \left\{ j : x_i^j = x_i \text{ for all } i \in S \cup H_k \right\}$ be the set of labels for those covariate vectors x^j that agree with the agent's covariate vector in entries $S \cup H_k$.

Further define

$$Z_k^n \equiv \frac{\sum_{j \in S_k} Y_j}{|S_k|}.$$

to be the prediction of the agent's type given his nonstandard covariates in H_k. The special case

$$Z_\varnothing^n \equiv \frac{1}{2^n} \sum_{j=1}^{2^n} Y_j$$

corresponds to disclosure of no nonstandard covariates. (Although $Z_\varnothing^n = Z_k^n$ for some k, we will use this more evocative notation when we want to highlight this special case.) To simplify notation, we subsequently drop the superscript n on both Z_k^n and $Z_\varnothing^n$.

A.2 Outline of Proof

Section A.3 proves the following result, which says that the difference between the highest prediction and the prediction $\mu \equiv \mathbb{E}[Z_\varnothing^n]$ vanishes as the number of covariates grows large.

Proposition A.1. $\lim_{n \to \infty} \mathbb{E}[\max_{1 \leq k \leq K_n} Z_k - \mu] = 0$.

Section A.4 strengthens this to the statement that the expected *absolute* difference between the highest prediction and μ vanishes.

Proposition A.2. $\lim_{n\to\infty} \mathbb{E}[\max_{1\leq k\leq K_n} |Z_k - \mu|] = 0$.

Section A.5 finally applies the above proposition to demonstrate the conclusion of the theorem, i.e., that

$$\lim_{n\to\infty} V(n, x_n) = \lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y)\right] - \mathbb{E}\left[u\left(Z^n_\varnothing, Y\right)\right] = 0$$

Thus in expectation the largest possible utility also vanishes. Since Assumption 1 implies that $V(n, x_n)$ does not depend on x_n, we simply write $V(n)$ going forward.

A.3 Proof of Proposition A.1

Statement of the proposition: $\lim_{n\to\infty} \mathbb{E}[\max_{1\leq k\leq K_n} Z_k - \mu] = 0$.

The quantity $\mathbb{E}[\max_{1\leq k\leq K_n} Z_k]$ is the expected first-order statistic of a sequence of non-i.i.d. variables $Z_1, \ldots, Z_{K_n}$. The proof is organized as follows. In Sections A.3.1 and A.3.2, we define i.i.d. variables Z^{iid}_k with the property that

$$\mathbb{E}\left[\max\{Z_1, \ldots, Z_{K_n}\}\right] \leq \mathbb{E}\left[\max\{Z^{iid}_1, \ldots, Z^{iid}_{K_n}\}\right]. \tag{A.1}$$

In Sections A.3.3 and A.3.4, we show that the RHS of the above display converges to μ as n grows large.

A.3.1 Replacing Z_k's with independent variables Z^{ind}_k

In general, disclosures k and k' may lead to posterior expectations Z_k and $Z_{k'}$ that are correlated due to the presence of the same Y_i's across the different sample averages. We first show that replacing these Z_k's with properly defined independent random variables weakly increases the value of context.

Definition A.1. *For each $1 \leq k \leq K_n$ define*

$$Z^{ind}_k = \frac{\sum_{j=1}^{|S_k|} Y^k_j}{|S_k|}$$

where $Y^k_j \sim_{iid} \pi_{x_S}$, so that each Z^{ind}_k has the same distribution as Z_k, but the vector $(Z^{ind}_1, \ldots, Z^{ind}_K)$ is mutually independent.

Lemma A.1. *Let* $V_n \equiv \mathbb{E}[\max Z_1, ..., Z_{K_n}]$ *and* $V_n^{ind} \equiv \mathbb{E}[\max\{Z_1^{ind}, ..., Z_{K_n}^{ind}\}]$. *Then* $V_n \leq V_n^{ind}$ *for all* $n \in \mathbb{Z}_+$.

Proof. Throughout we use $X \succeq Y$ to mean that the distribution of X first-order stochastically dominates the distribution of Y.

Sublemma 1. *Let* X_1,..., X_Q,W *be a sequence of real-valued random variables (not necessarily i.i.d.). Let* $a_1 > a_2 > ... > a_{Q-1} > a_Q > 0$ *be a sequence of positive constants. Further, let* $Y_1, ..., Y_Q$ *be i.i.d. random variables, independent of* $(X_1, \ldots, X_Q, W)$. *Define*

$$M_C = \max_{i \in \{1,...,Q\}} \{X_i + a_i Y_1\}$$
$$M_I = \max_{i \in \{1,...,Q\}} \{X_i + a_i Y_i\}$$

Then $M_I \succeq M_C$ *and* $\max\{M_I, W\} \succeq \max\{M_C, W\}$.

Proof. For $q \in \{1, ..., Q\}$ define:

$$M_C^q = \max\left\{\max_{i \in \{1,...,q-1\}} \{X_i + a_i Y_1\}, X_q + a_q Y_1\right\}$$
$$\widetilde{M}_C^q = \max\left\{\max_{i \in \{1,...,q-1\}} \{X_i + a_i Y_1\}, X_q + a_q Y_q\right\}$$

so that M_C^q is the maximum of the first q terms in M_C, and $\widetilde{M}_C^q$ replaces Y_1 in the q-th term of M_C^q with Y_q. We first demonstrate an analogue of the desired conclusions for M_C^q and $\widetilde{M}_C^q$.

Sublemma 2. $\widetilde{M}_C^q \succeq M_C^q$ *and* $\max\{\widetilde{M}_C^q, W\} \succeq \max\{M_C^q, W\}$.

Proof. Without loss of generality set $a_q = 1$. We'll first show that $\widetilde{M}_C^q \succeq M_C^q$. To establish first-order stochastic dominance, we need to show that for all $t \in \mathbb{R}$ it holds that $\mathbb{P}(M_C^q \leq t) - \mathbb{P}(\widetilde{M}_C^q \leq t) \geq 0$. For each $i \in \{1, ..., q-1\}$ define the event

$$B_i := \{X_q + Y_1 > X_i + a_i Y_1\} \equiv \left\{Y_1 < \frac{1}{a_i - 1}(X_q - X_i)\right\}.$$

Further let

$$B = \bigcap_{i=1}^{q} B_i = \left\{Y_1 < \min_{i \in \{1,...,q-1\}} \frac{1}{a_i - 1}(X_q - X_i)\right\}$$

be the event that $X_q + Y_1$ achieves the maximum among $\{X_i + a_i Y_1\}_{i=1}^q$. We'll show that the FOSD rankings in Sublemma 2 hold both on event B and also on its complement B^c.

Define

$$\widetilde{B} := \left(Y_q < \min_{i \in \{1,\ldots,q-1\}} \left\{ \frac{1}{a_i - 1}(X_q - X_i) \right\} \right)$$

to be the event that $X_q + Y_q$ achieves the maximum among $\{X_i + a_i Y_q\}_{i=1}^q$. Then

$$\begin{aligned}
\widetilde{M}_C^q | B &\succeq (X_q + Y_q)|B \\
&\stackrel{d}{=} X_q|B + Y_q && \text{since } Y_q \perp\!\!\!\perp (X_1, \ldots, X_q, Y_1) \\
&\succeq X_q|B + Y_q|\tilde{B} && \text{since } Y_q \succeq Y_q \mid \widetilde{B} \\
&\stackrel{d}{=} X_q|B + Y_1|B && \text{since } Y_1 \mid B \stackrel{d}{=} Y_q \mid \widetilde{B} \\
&\stackrel{d}{=} (X_q + Y_1)|B \stackrel{d}{=} M_C^q|B
\end{aligned}$$

Thus $\widetilde{M}_C^q|B \succeq M_C^q|B$.

Now consider the event B^c, on which $X_q + Y_1$ does not achieve the maximum among $\{X_i + a_i Y_1\}_{i=1}^q$. Then either $X_1 + Y_q \leq \max\{X_i + a_i Y_1\}_{i=1}^{q-1}$, in which case $\widetilde{M}_C^q = M_C^q$, or $X_1 + Y_q > \max\{X_i + a_i Y_1\}_{i=1}^{q-1}$, in which case $\widetilde{M}_C^q > M_C^q$. So

$$\widetilde{M}_C^q|B^c \succeq \max\{X_1 + a_1 Y_1, ..., X_{q-1} + a_{q-1} Y_1\}|B^c \stackrel{d}{=} M_C^q|B^c.$$

and hence $\widetilde{M}_C^q|B^c \succeq M_C^q|B^c$.

Now we show that $\max\{\widetilde{M}_C^q, W\} \succeq \max\{M_C^q, W\}$. For any realization w of W, let X_i^w denote the conditional random variable $X_i|W = w$. Define $M_C^{q,w}$ and $\widetilde{M}_C^{q,w}$ identically to M_C^q and $\widetilde{M}_C^q$, replacing each X_i by X_i^w. Then by independence of W and $(Y_1, \ldots, Y_q)$, the distribution of $\max\{M_C^{q,w}, w\}$ is identical to that of $\max\{M_C^q, W\}|W = w$, and the distribution of $\max\{\widetilde{M}_C^{q,w}, w\}$ is identical to that of $\max\{\widetilde{M}_C^q, W\}|(W = w)$.

Applying the first part of this sublemma to $M_C^{q,w}$ and $\widetilde{M}_C^{q,w}$, we conclude that $M_I^{q,w} \succeq M_C^{q,w}$. Since $\max\{., w\}$ is an increasing convex function, it preserves the first-order stochastic dominance relation and hence $\max\{\widetilde{M}_C^q, W\}|(W = w) \succeq \max\{M_C^q, W\}|(W = w)$. This argument holds pointwise for all w so $\max\{\widetilde{M}_C^q, W\} \succeq \max\{M_C^q, W\}$ as desired. □

We now complete the proof that $\max\{M_C, W\} \succeq \max\{M_I, W\}$. From similar (omitted) arguments it follows that $M_I \succeq M_C$. For each $q \in \{1, \ldots, Q-1\}$ define

$$\widehat{M}_C^q = \max\left\{ \max\{X_i + a_i Y_1\}_{i=1}^q, \max\{X_i + a_i Y_i\}_{i=q+1}^Q, W \right\}$$

observing that $\max\{M_I, W\} = \widehat{M}_C^1$ and that $\widehat{M}_C^Q \succeq \max\{M_C, W\}$ (by Sublemma 2). Moreover for each $q \in \{1, \ldots, Q-1\}$, we have $\widehat{M}_C^q = \max\{M_C^q, W^q\}$ and $\widehat{M}_C^{q-1} = \max\{\widetilde{M}_C^q, W^q\}$

where $W^q = \max\left\{\max\{X_i + a_i Y_i\}_{i=q+1}^{Q}, W\right\}$ is independent of $(Y_1, \ldots, Y_{q-1})$. So applying Sublemma 2, $\widehat{M}_C^{q-1} \succeq \widehat{M}_C^q$ as desired. □

Finally, we use Sublemma 1 to establish Lemma A.1, i.e., the expected value of context weakly increases if we make the Y's within different disclosures independent. We will prove this iteratively. For arbitrary $n \in \mathbb{N}$, define the random variable

$$M = \max\{Z_1, \ldots, Z_{K_n}\} = \max\left\{\frac{\sum_{j\in S_1} Y_j}{|S_1|}, \ldots, \frac{\sum_{j\in S_{K_n}} Y_j}{|S_{K_n}|}\right\}.$$

Fix any Y_i. We will show that replacing Y_i across different sample averages with independent copies of this random variable leads to a FOSD increase in the distribution of M.

Let $I = \{k : i \in S_k\}$ be the set of indices of sample averages which contain Y_i. Then we can rewrite the previous display as

$$\max\left\{\max_{k\in I} \frac{\sum_{j\in S_k} Y_j}{|S_k|}, \max_{k\notin I} \frac{\sum_{j\in S_k} Y_j}{|S_k|}\right\}$$

or

$$\max\left\{\max_{k\in I}\left\{X_k + \frac{1}{|S_k|}Y_i\right\}, W\right\} \tag{A.2}$$

where $X_k \equiv \frac{1}{|S_k|}\sum_{j\in S_k, j\neq i} Y_j$ for each $k \in I$, and $W \equiv \max_{k\notin I} \frac{\sum_{j\in S_k} Y_j}{|S_k|}$. Because $(Y_1, \ldots, Y_{K_n})$ are mutually independent, Y_i is independent of each X_k and W. So applying Lemma 1, the random variable in (A.2) has a distribution that is first-order stochastically dominated by the distribution of

$$\max\left\{\max_{k\in I}\left\{X_k + \frac{1}{|S_k|}Y_i^k\right\}, W\right\}$$

as desired. Since Y_i is arbitrary, this concludes the proof. □

A.3.2 Replacing Z_k^{ind} with i.i.d. Variables Z_k^{iid}

The variables $Z_1^{ind}, \ldots, Z_{K_n}^{ind}$ are sample averages of unequal sizes ranging between 2^{n-h_n} and 2^n elements. We next show that replacing each of these variables with a sample average of 2^{n-h_n} elements (the smallest size) weakly increases the value of context.

Definition A.2. *For each $1 \leq k \leq K_n$ define*

$$Z_k^{iid} = \frac{\sum_{j=1}^{2^{n-h_n}} Y_j^k}{2^{n-h_n}}$$

to be the analogue of Z_k^{ind} with 2^{n-h_n} elements instead of $|S_k| \geq 2^{n-h_n}$, so that the variables $Z_1^{iid}, \ldots, Z_{K_n}^{iid}$ are iid.

Lemma A.2. *Let* $V_n^{iid} \equiv \mathbb{E}\left[\max\{Z_1^{iid}, \ldots, Z_{K_n}^{iid}\}\right]$. *Then* $V_n^{ind} \leq V_n^{iid}$ *for all* $n \in \mathbb{Z}_+$.

Proof. We use the following result.

Sublemma 3. *Suppose* $Y_1, Y_2, \ldots, Y_n$ *are independent and identically distributed random variables, and define* $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ *to be their sample average. Let* $n' < n$ *and define* $\overline{Y}_{n'} = \frac{1}{n'}\sum_{i=1}^{n'} Y_i$. *Then the distribution of* $\overline{Y}_{n'}$ *is a mean-preserving spread of the distribution of* $\overline{Y}_n$.

Proof. First observe that $\mathbb{E}[Y_j \mid \overline{Y}_n] = \overline{Y}_n$ for any $j = 1, \ldots, n$, since

$$\overline{Y}_n = \mathbb{E}[\overline{Y}_n \mid \overline{Y}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i \mid \overline{Y}_n] = \mathbb{E}[Y_j \mid \overline{Y}_n]$$

where the final equality follows by assumption that the Y_i's are iid. Then $\mathbb{E}[\overline{Y}_{n'} \mid \overline{Y}_n] = \frac{1}{n'}\sum_{i=1}^{n'} \mathbb{E}[Y_i \mid \overline{Y}_n] = \frac{1}{n'}\sum_{i=1}^{n'} \overline{Y}_n = \overline{Y}_n$ and the distribution of $\overline{Y}_{n'}$ is a mean-preserving spread of the distribution of $\overline{Y}_n$ as desired. □

This lemma implies that each Z_k^{iid} second-order stochastically dominates Z_k^{ind} (since $|S_k| \geq 2^{n-h_n}$ for all k). The result then follows by Jensen's inequality, since the entries of $(Z_1^{ind}, \ldots, Z_K^{ind})$ are (by construction) independent and maximum is a convex function. □

A.3.3 Asymptotic Normality

Lemma A.3. *Let* $V_n^N \equiv \mathbb{E}\left[\max\{Z_1^N, \ldots, Z_{K_n}^N\}\right]$ *where* $Z_k^N \sim N\left(\mu, \frac{1}{2^{n-h_n}}\right)$. *Then*

$$\lim_{n\to\infty} |V_n^{iid} - V_n^N| = 0.$$

Proof. Without loss of generality, let $Var(Y_j^k) = 1$.[26] First observe that

$$\sqrt{2^{n-h_n}} \cdot V_n^{iid} = \mathbb{E}\left[\max\{\widetilde{Z}_1^{iid}, \ldots, \widetilde{Z}_{K_n}^{iid}\}\right]$$

where each $\widetilde{Z}_k^{iid} = \frac{1}{\sqrt{2^{n-h_n}}}\sum_{i=1}^{2^{n-h_n}} Y_i^k$. Similarly we can write

$$\sqrt{2^{n-h_n}} \cdot V_n^N = \mathbb{E}\left[\max\{\widetilde{Z}_1^N, \ldots, \widetilde{Z}_{K_n}^N\}\right]$$

where each $\widetilde{Z}^N \sim_{iid} N(\mu, 1)$. When the assumptions for Corollary 2.1 from Chernozhukov et al. (2013) are met (to be verified momentarily), we can conclude that

$$\rho\left(\max\{\widetilde{Z}_1^{iid}, \ldots, \widetilde{Z}_{K_n}^{iid}\}, \max\{\widetilde{Z}_1^N, \ldots, \widetilde{Z}_{K_n}^N\}\right) \to 0$$

[26] If $Var(Y_j^k) = 0$, the statement of Theorem 1 holds trivially.

where ρ denotes Kolmogorov distance. Thus also

$$\rho(M_n^{iid}, M_n^N) \to 0 \tag{A.3}$$

where $M_n^{iid} = \frac{1}{\sqrt{2^{n-h_n}}} \max\{\widetilde{Z}_1^{iid}, \ldots, \widetilde{Z}_{K_n}^{iid}\}$ and $M_n^N = \frac{1}{\sqrt{2^{n-h_n}}} \max\{\widetilde{Z}_1^N, \ldots, \widetilde{Z}_{K_n}^N\}$.

By assumption, each Y_i^k is supported on $[-\overline{y}, \overline{y}]$ for some finite $\overline{y}$. This implies $|M_n^{iid}| \leq \overline{y}$ for all n, so the sequence $(M_n^{iid})_n$ is uniformly integrable. The convergence in (A.3) thus implies $\lim_{n\to\infty} \left|\mathbb{E}\left[M_n^{iid}\right] - \mathbb{E}\left[M_n^N\right]\right| = \lim_{n\to\infty} |V_n^{iid} - V_n^N| = 0$ as desired.

It remains to verify that the conditions of Corollary 2.1 from Chernozhukov et al. (2013) are met. This follows from the assumption that Y_j^k's are uniformly bounded, and the observation that

$$\frac{\left(\log(K_n \cdot 2^{n-h_n})\right)^7}{2^{(1-c)(n-h_n)}} \xrightarrow{n\to\infty} 0$$

for any $c \in (0,1)$, since $K_n = \sum_{j=0}^{h_n} \binom{n}{j} \leq 2^n$ by the Binomial Theorem and $\alpha_h < 1$. □

A.3.4 Upper Bound for Expected Maximum of Gaussians

Finally by Berman (1964), which provides an asymptotic upper bound for the expected maximum of independent Gaussian random variables

$$V_n^N \leq \frac{1}{\sqrt{2^{n-h_n}}} \cdot \sqrt{2\log(K_n)} \leq \frac{1}{\sqrt{2^{n(1-\alpha_h)}}} \cdot \sqrt{2n}$$

where the final expression converges to zero as $n \to \infty$ by assumption that $\alpha_h < 1$. Since clearly $\mathbb{E}[\max_{1\leq k\leq K_n} Z_k - \mu] \geq 0$, this concludes the proof of Proposition A.1.

A.4 Proof of Proposition A.2

Statement of the proposition: $\lim_{n\to\infty} \mathbb{E}[\max_{1\leq k\leq K_n} |Z_k - \mu|] = 0$.

In an abuse of notation, let $Z_k \equiv Z_k - \mu$ denote de-meaned sample average. Note that the de-meaned average $Z_k \in [-2\overline{y}, 2\overline{y}]$ is uniformly bounded. By rewriting the max within the expectation we obtain

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k|\right] &= \mathbb{E}\left[\max\left\{\max_{1\leq k\leq K_n} Z_k, -\min_{1\leq k\leq K_n} Z_k\right\}\right] \\
&\leq \mathbb{E}\left[\max\left\{\max_{1\leq k\leq K_n} \{Z_k\}, 0\right\}\right] + \mathbb{E}\left[\max\left\{-\min_{1\leq k\leq K_n} \{Z_k\}, 0\right\}\right]
\end{aligned}$$

We will show that each term of this final expression converges to zero. Observe that

$$\mathbb{E}\left[\max\left\{\max_{1\leq k\leq K_n}\{Z_k\},0\right\}\right] = \mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k \geq 0\right)\cdot\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k \geq 0\right] \quad \text{(A.4)}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] &= \mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k \geq 0\right)\cdot\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k \geq 0\right] \\ &\quad + \mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k < 0\right)\cdot\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k < 0\right]\end{aligned}$$

so

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k \geq 0\right)\cdot\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k \geq 0\right] = \\ &\qquad = \mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] - \mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k < 0\right)\cdot\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k < 0\right]\end{aligned}$$

From Lemma A.1, $\lim_{n\to\infty}\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] = 0$. Moreover, we showed in Section A.3.1 that the distribution of $(Z_1^{ind},\ldots,Z_{K_n}^{ind})$ first-order-stochastically-dominates that of $(Z_1,\ldots,Z_{K_n})$, so

$$\mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k < 0\right) \leq \mathbb{P}\left(\max_{1\leq k\leq K_n} Z_k^{ind} < 0\right) \leq \prod_{1\leq k\leq K_n}\mathbb{P}(Z_k^{ind} < 0)$$

which converges to zero as n grows large. Indeed, because every Z_k^{ind} is a sample mean of at least 2^{n-h_n} zero-mean draws, the Chebyshev inequality gives a uniform constant $\delta > 0$ such that $\mathbb{P}(Z_k^{ind} < 0) \leq 1-\delta$ for all large n. Since K_n grows at least exponentially in n, $(1-\delta)^{K_n} \leq \exp(-\delta 2^{cn}) \to 0$, which completes the argument. Finally,

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k \mid \max_{1\leq k\leq K_n} Z_k < 0\right] \in [-2\overline{y}, 2\overline{y}] \quad \text{(A.5)}$$

uniformly across n. Putting together (A.4) - (A.5) we have that

$$\lim_{n\to\infty}\mathbb{E}\left[\max\left\{\max_{1\leq k\leq K_n}\{Z_k\},0\right\}\right] = 0$$

as desired. The argument that $\lim_{n\to\infty}\mathbb{E}\left[\max\left\{-\min_{1\leq k\leq K_n}\{Z_k\},0\right\}\right] = 0$ follows identically, observing that Proposition A.1 is satisfied for $\widetilde{Y} \equiv -Y$, and that

$$-\min_{1\leq k\leq K_n} Z_k = \max_{1\leq k\leq K_n} -\frac{\sum_{j\in S_k} Y_j}{|S_k|} = \max_{1\leq k\leq K_n}\frac{\sum_{j\in S_k}\widetilde{Y}_j}{|S_k|}.$$

A.5 Concluding the proof of Theorem 1

Recall that $Z^n_\varnothing \equiv \frac{1}{2^n}\sum_{j=1}^{2^n} Y_j$ denotes the (random) posterior expectation when the agent chooses not to disclose any nonstandard covariates. Clearly $V(n) \geq 0$ (since the agent can always choose to disclose nothing). Also

$$V(n) = \mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y)\right] - \mathbb{E}\left[u(Z^n_\varnothing, Y)\right] \leq \mathbb{E}\left[\max_{1\leq k\leq K_n} |u(Z_k, Y) - u(Z^n_\varnothing, Y)|\right]$$

Each absolute difference $|u(Z_k, Y) - u(Z^n_\varnothing, Y)|$ can be bounded from above using the triangle inequality

$$|u(Z_k, Y) - u(Z^n_\varnothing, Y)| \leq |u(Z_k, Y) - u(\mu, Y)| + |u(\mu, Y) - u(Z^n_\varnothing, Y)| \tag{A.6}$$

Since u is by assumption Lipschitz continuous in the first argument, there is a constant B such that $|u(z_k, y) - u(\mu, y)| \leq B|z_k - \mu|$ and $|u(\mu, y) - u(z_\varnothing, y)| \leq B|z_\varnothing - \mu|$ for any realizations z_k and $z_\varnothing$ of Z_k and $Z^n_\varnothing$. Combining these inequalities, we get

$$V(n) \leq B\left(\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k - \mu|\right] + \mathbb{E}\left[|Z^n_\varnothing - \mu|\right]\right)$$

First, note that $\mathbb{E}[Z^n_\varnothing] = \mu$. Moreover, by assumption that each Y is uniformly bounded above and below, the sequence $(Z^n_\varnothing)$ is uniformly integrable. It follows from the Law of Large Numbers that $\lim_{n\to\infty} \mathbb{E}[|Z^n_\varnothing - \mu|] = 0$. Finally, $\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k - \mu|\right] = 0$ follows directly from Lemma A.2. So the RHS of A.6 converges to zero, implying $V(n) \to 0$.

A.6 Proof of Theorem 2

Throughout the proof we set $s = 0$, $\mu = 0$ and $\sigma^2 = \mathbb{E}[Y^2] = 1$ without loss of generality.

(a) Let $A_n \subseteq \{1, \ldots, 2^n\}$ index those 2^{n-b_n} covariate vectors that agree with the agent's covariate vector for all covariates in A. Then the algorithmic institution's posterior expectation is the sample average $Z^n_A = \frac{1}{2^{n-b_n}}\sum_{j\in A_n} Y_j$. We will first show that

$$\begin{aligned}\Delta(n) &\equiv \mathbb{E}[\phi(Z^n_A)] - \mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k)\right] \\ &= \mathbb{E}[\phi(Z^n_A) - \phi(0)] - \mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] > 0\end{aligned}$$

for large enough n if $\alpha_b > \frac{1+\alpha_h}{2}$. Then we will show that $\Delta(n) < 0$ for large enough n if $\alpha_b < \frac{1+\alpha_h}{2}$.

We start by analyzing the first difference $\mathbb{E}[\phi(Z_A^n) - \phi(0)]$. Using Taylor's expansion we get

$$\mathbb{E}[\phi(Z_A^n) - \phi(0)] = \mathbb{E}[\phi'(0)Z_A^n] + \mathbb{E}\left[\frac{\phi''(\tilde{Z})}{2}(Z_A^n)^2\right]$$

for some $\tilde{Z} \in [0, Z_A^n]$. Note that $\mathbb{E}[Z_A^n] = \mathbb{E}[Y] = 0$. Moreover, $\phi''(\tilde{Z}) \geq M_2 > 0$, where $M_2 = \min_{\hat{y}\in[-\overline{y},\overline{y}]} |\phi''(\hat{y})|$. Thus

$$\mathbb{E}[\phi(Z_A^n) - \phi(0)] \geq \frac{M_2}{2}\mathbb{E}[(Z_A^n)^2] = \frac{M_2}{2 \cdot 2^{(1-\alpha_b)n}} \tag{A.7}$$

Next turn to $\mathbb{E}[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)]$. For each term inside the maximum we have that

$$\phi(Z_k) - \phi(0) \leq M_1|Z_k| \tag{A.8}$$

where $M_1 = \max_{\hat{y}\in[-\overline{y},\overline{y}]} |\phi'(\hat{y})|$. Thus $\mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] \leq M_1\mathbb{E}\left[\max\{|Z_k|\}\right]$.

By Proposition A.2 and following analogous steps to the proof of Theorem 1 we can show $\mathbb{E}[\max\{|Z_k|\}] \leq \frac{\sqrt{2}}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}$. Thus

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] \leq M_1\sqrt{2}\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}$$

Combining the bounds from steps 1 and 2 we get

$$\Delta(n) \geq M_2\frac{1}{2 \cdot 2^{(1-\alpha_b)n}} - M_1\sqrt{2}\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}$$

The RHS is positive for all large n if and only if $\frac{2^{\frac{(1-\alpha_h)}{2}n}}{2^{(1-\alpha_b)n}} \xrightarrow{n\to\infty} \infty$, since $\sqrt{\log(2K_n)}$ has sub-exponential but non-constant asymptotics. This condition is satisfied if $\alpha_b > \frac{1+\alpha_h}{2}$.

Now we demonstrate that when $\alpha_b < \frac{1+\alpha_h}{2}$ we have $\Delta(n) < 0$ for all large enough n. Suppose $\phi'(\mu) \neq 0$. Using Taylor's expansion we get

$$\mathbb{E}[\phi(Z_A^n) - \phi(0)] \leq c_3\mathbb{E}[(Z_A^n)^2] = \frac{c_3}{2^{(1-\alpha_b)n}}$$

where, as before, we used $\mathbb{E}[Z_A^n] = 0$ and the fact that By Assumption 2 the curvature ϕ'' is bounded from above, so $\phi''(.) < 2M_3$ for some $M_3 > 0$. Expanding every term inside $\max_k \phi(Z_k)$ in the same way we get:

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] = \mathbb{E}\left[\max_{1\leq k\leq K_n} \phi'(0)Z_k + \frac{\phi''(\tilde{Z}_k)}{2}(Z_k)^2\right]$$

for some $\tilde{Z}_k \in [0, Z_k]$. Since $\phi'(0) \neq 0$ and $\phi''(.) > 0$ we get

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] \tag{A.9}$$

$$> |\phi'(0)|\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] \tag{A.10}$$

$$\geq |\phi'(0)|C\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(K(n))} \tag{A.11}$$

for some $C > 0$ and some sequence $K(n)$. The last inequality follows from lemma A.4, which we prove below. In the lemma we also establish that $K(n)$ has sub-exponential asymptotics. Combining the inequalities A.6— A.11 we get

$$\Delta(n) = \mathbb{E}[\phi(Z_A^n)] - \mathbb{E}[\max_{1\leq k\leq K_n} \phi(Z_k)] \tag{A.12}$$

$$< \frac{M_3}{2^{(1-\alpha_b)n}} - |\phi'(0)|C\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(K(n))} \tag{A.13}$$

The RHS of A.13 is asymptotically negative if and only if $\alpha_b < \frac{1+\alpha_h}{2}$.

Lemma A.4. *For some $c, C > 0$ and all sufficiently large n:*

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] \geq C\frac{\sqrt{\log(cK_n)}}{\sqrt{2^{(1-\alpha_h)n}}}$$

Proof. We will show the result in two steps. First, we will construct a specific lower bound on $\mathbb{E}[\max_k Z_k]$. We will then show that this lower bound can be approximated by an maximum of Gaussian random variables by applying Corollary 2.1 from Chernozhukov et al. (2013) to the constructed lower bound. This will allow us to use a known lower bound for the expected maximum of Gaussian random variables, which will yield the result.

Fix n and some $K \in \mathbb{N}$, and choose any $H_1, ..., H_K \in \mathcal{H}_n$ such that each $|H_k| = \lfloor \alpha_h n \rfloor$. Then

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] \geq \mathbb{E}\left[\max_{1\leq k\leq K} \hat{y}(H_k)\right]$$

since the right-hand side restricts attention to sets H_k that reveal exactly $\lfloor \alpha_h n \rfloor$ covariates. We will show that $\mathbb{E}\left[\max_{1\leq k\leq K} \hat{y}(H_k)\right] \geq C\frac{\sqrt{\log(cK_n)}}{\sqrt{2^{(1-\alpha_h)n}}}$.

Let $m = \lfloor (1-\alpha_h)n \rfloor + 1$, the number of covariates that remain hidden after each disclosure. Define $\hat{y} := \{\hat{y}(H_k)\}_{k=1}^K$, the random K-dimensional vector whose kth component is the prediction corresponding to set H_k. To use the Gaussian approximation, we will now show

that for any K there always exists n large enough and $H_1, ..., H_K \in \mathcal{H}_n$ such that $|H_k| = \lfloor \alpha_h n \rfloor$ and

$$\hat{y} = \frac{1}{2^m} \sum_{1 \leq j \leq 2^m} X_j$$

for some independent random K-vectors $\{X_j\}_{j=1}^{2^m}$.

We start by demonstrating the result for $K = 3$, and then show how the proof generalizes to arbitrary K. Suppose $K = 3$. We will find three feasible disclosures, for which the vector of posteriors can be written as

$$\hat{y} = \frac{1}{2^m} \sum_{1 \leq j \leq 2^m} X_j$$

where the random 3-vectors X_j are independent of each other. Assume n is large enough that $\lfloor \alpha_h n \rfloor \geq 2$ (which always holds for sufficiently large n since $\alpha_h > 0$). Consider the following three disclosure sets: 1) I_1 : hide covariates $\{1, ..., m\}$, 2) I_2 : hide covariates $\{1, ..., m-1, m+1\}$, 3) I_3 : hide covariates $\{1, ..., m-1, m+2\}$.

Let $\mathcal{Y}_j = \{Y_{x'} | x' \in \Pi_{A_i}(x)\}$ be the set of types Y_x consistent with disclosure I_i. These sets have the same size $|\mathcal{Y}_j| = 2^m$. Moreover, the sets $\mathcal{Y}_1, \mathcal{Y}_2$ and $\mathcal{Y}_3$ have a common intersection: $\mathcal{Y}_i \cap \mathcal{Y}_j = \mathcal{Y}_1 \cap \mathcal{Y}_2 \cap \mathcal{Y}_3 =: C$. Indeed, define a disclosure I_0 as follows: hide covariates $\{1, ..., m-1\}$. Then, assuming without loss $x = (1, ..., 1)$, we have

$$\Pi_{I_1}(x) = \Pi_{I_1}((1, ..., 1)) = \Pi_{I_0}((1, ..., 1)) \cup \Pi_{I_0}(1, ..., 1, \overset{m}{0}, 1, ..., 1)$$

Similarly

$$\Pi_{I_2}(x) = \Pi_{I_0}((1, ..., 1)) \cup \Pi_{I_0}(1, ..., 1, \overset{m+1}{0}, 1, ..., 1)$$

$$\Pi_{I_3}(x) = \Pi_{I_0}((1, ..., 1)) \cup \Pi_{I_0}(1, ..., 1, \overset{m+2}{0}, 1, ..., 1)$$

So $\Pi_{I_i} \cap \Pi_{I_j} = \Pi_{I_0}(x)$, as desired. It also follows that $|C| = |\Pi_{I_0}(x)| = 2^{m-1}$. Enumerate the elements c_i of C with index $i \in \{1, ..., 2^{m-1}\}$. Further, enumerate the elements of each $\mathcal{Y}_j \setminus C$ with index $i \in \{2^{m-1} + 1, ..., 2^m\}$. Then, construct vectors X_i as follows: For $i \in \{1, ..., 2^{m-1}\}$ let $X_i = c_i \mathbf{1}_3$ for $c_i \in C$. (Note that by construction $c_i \neq c_j$ for $i \neq j$.) And for $i \in \{2^{m-1} + 1, 2^m\}$ let $(X_i)^j = y_i^j$ where y_i^j is the i'th element of $\mathcal{Y}_j \setminus C$. The vectors X_j are independent by construction.

To generalize the construction of X_j to an arbitrary K, suppose n is large enough to satisfy $\lfloor \alpha_h n \rfloor \geq K - 1$. As before, such n always exist as long as $\alpha_h \neq 0$. The rest of the construction proceeds similarly to $K = 3$ with the sets I_k defined accordingly.

Let $K(n)$ be the largest number K, for which $\lfloor \alpha_h n \rfloor \geq K-1$. By the above construction, $K(n)$ is an increasing sequence. Moreover, $K(n) \leq K_n$. By construction, $\hat{y} = \frac{1}{2^m}\sum_{1\leq j\leq 2^m} X_j$. Since the vectors X_j are independent, we can apply Corollary 2.1 from Chernozhukov et al. (2013) to the quantity $\max_k \hat{y}_k$ directly. As above, let $\mathcal{Y}_k$ be the set of types that are consistent with the k'th disclosure out of $K(n)$. Then

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] \geq \mathbb{E}\left[\max_{1\leq k\leq K(n)} \hat{y}_k\right] = \mathbb{E}\left[\max_{1\leq k\leq K(n)} \left(\frac{1}{2^m}\sum_{1\leq j\leq 2^m} X_j\right)_k\right] \tag{A.14}$$

where, as before, $m = \lfloor (1-\alpha_h)n \rfloor + 1$. By construction $X_j = c_j \mathbf{1}_{K(n)}$ for all $j \leq 2^{m-1}$. Taking this into account, the right hand side of (A.14) can be rewritten as

$$\mathbb{E}\left[\frac{1}{2}\frac{\sum_{Y_x\in C} Y_x}{2^{m-1}}\right] + \mathbb{E}\left[\max_{1\leq k\leq K(n)} \frac{1}{2}\frac{\sum_{Y_x\in \mathcal{Y}_k\setminus C} Y_x}{2^{m-1}}\right]$$

The first term equals to zero. Since in the second term all Y_x are independent, Corollary 2.1 from Chernozhukov et al. (2013) applies directly. Hence, following the steps from the proof of Theorem 1 and applying the strict lower bound on the expected maximum of gaussian random variables with the appropriate constant we get

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} Z_k\right] > const \cdot \sqrt{\frac{\log(const \cdot K(n))}{2^{(1-\alpha_h)n}}}$$

which proves the result. □

(b) Since $-\phi$ is convex, the above arguments apply to show that $-\frac{M_3}{2^{(1-\alpha_b)n}} \leq \mathbb{E}[\phi(Z_A^n) - \phi(0)] \leq -\frac{M_2}{2\cdot 2^{(1-\alpha_b)n}}$ and

$$\begin{aligned}\mathbb{E}\left[\min_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] &= -\mathbb{E}\left[\max_{1\leq k\leq K_n} -\phi(Z_k) - (-\phi(0))\right] \\ &\geq -M_1\sqrt{2}\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}\end{aligned}$$

for some $c_4 > 0$. The desired conclusion follows. The converse for the case $\alpha_b < \frac{1+\alpha_h}{2}$ follows analogously to part a) with

$$\mathbb{E}\left[\min_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] \leq -|\phi'(0)|c_4\sqrt{2}\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}.$$

Online appendix to the paper

The Value of Context: Human versus Algorithmic Institutions

Andrei Iakovlev Annie Liang

August 26, 2025

P Supplementary Material to Section 4

P.1 Result Extending Theorem 2 Part (a)

Consider a model in which the evaluator chooses an action a given the realization of the agent's covariates, and the evaluator and agent share the payoff function $-(a-y)^2$. The following result shows that the conclusion of Part (a) of Theorem 2 extends for non-binary types y.

Proposition P.1. *There exists an N sufficiently large such that the agent robustly prefers the algorithmic institution for all $n \geq N$.*

Proof. Throughout the proof set $s = 0$, $\mathbb{E}[Y] = 0$ and $\sigma^2 = \mathbb{E}(Y^2) = 1$ without loss. We will show that

$$\begin{aligned}\mathbb{E}[u(Z_A^n, Y)] - \mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y)\right] \\ = \mathbb{E}[u(Z_A^n, Y) - u(0, Y)] - \mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y) - u(0, Y)\right] > 0\end{aligned}$$

for large enough n.

Let $x_A = (x_i)_{i\in A}$ denote the covariates that the algorithmic institution observes, and as before let $Z_A^n = \mathbb{E}[Y \mid x_A]$ denote algorithmic institution's (random) posterior expectation. The optimal action choice $a = Z_A^n$ yields expected payoff $-\text{Var}(Y \mid x_A)$. By the Law of Total Variance, $\mathbb{E}[-\text{Var}(Y \mid x_A)] = \text{Var}(Z_A^n) - \text{Var}(Y)$. Since additionally $\mathbb{E}[u(0, Y)] = -\text{Var}(Y)$, we obtain

$$\mathbb{E}[u(Z_A^n, Y) - u(0, Y)] = \mathbb{E}\left[(Z_A^n)^2\right] = \frac{1}{2^{(1-\alpha_b)n}}.$$

Now turn to $\mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y) - u(0, Y)\right]$. By Lipschitz continuity of u, there is a constant B such that $u(z_k, y) - u(0, y) \leq B|z_k|$ holds pointwise for each realization (z_k, y) of (Z_k, Y). So

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} u(Z_k, Y) - u(0, Y)\right] \leq c_2\mathbb{E}[\max\{|Z_k|\}]$$

The remainder of the proof proceeds identically to the proof of Theorem 2. □

P.2 Proof of Corollary 1

Following the proof of Theorem 2 we will prove the corollary assuming ϕ is strictly convex (the case of concave ϕ follows from similar computations). Suppose $\alpha_b > \frac{1+\alpha_h}{2}$. We need to show that for all $n \geq \max\{N_\phi, N_f\}$ we have that

$$\Delta(n) = \mathbb{E}\left[U_x^f(A)\right] - \mathbb{E}\left[\max_{H\in\mathcal{H}_n} U_x^f(H)\right] \geq 0$$

We start by establishing that for all $N_f \geq n$ we have

$$\Delta(n) \geq M_2\frac{1}{2\cdot 2^{(1-\alpha_b)n}} - M_1\sqrt{2}\frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{\log(2K_n)}$$

By Lipschitz continuity of ϕ we have

$$V(n, x_n) \leq M_1\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k - \mu|\right]$$

By construction, for all $n \geq N_f$ we obtain

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} |Z_k - \mu|\right] \leq \frac{1}{\sqrt{2^{(1-\alpha_h)n}}}\sqrt{2\log(2K_n)}$$

Further, convexity of ϕ implies that for all n

$$\mathbb{E}\left[U_x^f(A)\right] \geq M_2\mathbb{E}[Z_A^2] = \frac{M_2}{2^{(1-\alpha_b)n}}$$

which gives the desired lower bound on $\Delta(n)$.

To finish the proof, we show that $\Delta(n) > 0$ for all $n \geq N_\phi$. Since $2K_n \leq 2^{n+1} < e^{n+1}$, we have $\Delta(n) > 0$ if

$$\left(\alpha_b - \frac{\alpha_h + 1}{2}\right)n - \frac{1}{2}\log_2(n+1) > \log_2\left(2\sqrt{2}\frac{M_1}{M_2}\right)$$

which proves the result.

Q Proofs for Results in Sections 5

Q.1 Proof of Corollary 2

We follow the notation and setup of the proof of Theorem 1. Fix any realization $x_S = (x_1, \ldots, x_s)$ of the standard covariates. As in the proof of Theorem 1, there are 2^n covariate vectors $x_n \in \mathcal{X}_n$ with positive probability conditional on x_S. Index these by $j = 1, \ldots, 2^n$, and define

$$Y_j^{x_S} \equiv f\left(x_n^j\right)$$

to be the random type given covariate vector x_n^j. For each covariate vector x_n and disclosure set $H_k \in \mathcal{H}_n$, let S_k again denote those covariate vectors that agree with x_n on entries $S \cup H_k$, and define

$$Z_k^{x_S} = \frac{\sum_{j \in S_k} Y_j^{x_S}}{|S_k|}$$

to be the prediction of the agent's type given his covariates in $S \cup H_k$. Different from the proof of Theorem 1, there are now $\overline{K}_n = \sum_{j=0}^{h_n} \binom{n}{j} 2^j$ unique sets S_k (ranging over not only the different possible sets of covariates to disclose but also their values). By the Binomial Theorem,

$$\sum_{j=0}^{h_n} \binom{n}{j} 2^j \leq \sum_{j=0}^{n} \binom{n}{j} 2^j = 3^n.$$

Following the proof of Lemma A.1, we obtain that

$$\mathbb{E}\left(\max_{1 \leq k \leq \overline{K}_n} |Z_k^{x_S} - \mu|\right) \leq \frac{1}{\sqrt{2^{n-h_n}}} \sqrt{2 \log(2\overline{K}_n)} \leq \frac{1}{\sqrt{2^{n(1-\alpha_h)}}} \sqrt{2 \log(2 \cdot 3^n)}$$

which again converges to zero by assumption that $\alpha_h < 1$. Finally observe that

$$\begin{aligned} \mathbb{E}\left[\max_{x_S \in \{0,1\}^s} \left(\max_{1 \leq k \leq \overline{K}_n} |Z_k^{x_S} - \mu|\right)\right] &\leq \mathbb{E}\left[\sum_{x_S \in \{0,1\}^s} \max_{1 \leq k \leq \overline{K}_n} |Z_k^{x_S} - \mu|\right] \\ &= \sum_{x_S \in \{0,1\}^s} \mathbb{E}\left[\max_{1 \leq k \leq \overline{K}_n} |Z_k^{x_S} - \mu|\right]. \end{aligned}$$

Since each $\mathbb{E}\left[\max_{1 \leq k \leq \overline{K}_n} |Z_k^{x_S} - \mu|\right] \to 0$ as $n \to \infty$, the RHS converges to zero. We thus obtain the analogue of Lemma A.2 for the expected maximum value of context, and the remainder of the proof proceeds identically to Theorem 1.

Q.2 Proof of Proposition 1

Throughout this proof, we set $s = 0$.

Let (σ^*, η^*) denote a typical PBE, where σ^* is the Sender's disclosure strategy and η^* is the Receiver's evaluation function. Fixing any such equilibrium, we use $Z_{\eta^*}(d)$ to denote the Receiver's posterior expectation given disclosure d. We first prove that at least one pure-strategy equilibrium always exists.

Proposition Q.1. *For every n and f there exists a pure-strategy f-context equilibrium.*

Proof. Consider a candidate equilibrium (σ^*, η^*), where $\sigma^*(\mathbf{x}_n) = \varnothing$ for all $\mathbf{x}_n \in \mathcal{X}_n$ (which is clearly a feasible disclosure for all agents). The Receiver's beliefs at disclosure $\varnothing$ are pinned down by Bayes' rule. For any other disclosure $d \neq \varnothing$, we construct the agent's out-of-equilibrium beliefs such that $\phi(Z_{\eta^*}(\varnothing)) \geq \phi(Z_{\eta^*}(d))$. This is always possible, for example by setting $Z_{\eta^*}(\varnothing) = Z_{\eta^*}(d)$ for every d. Then by construction reporting $\varnothing$ is a best response for any $\mathbf{x}_n$, so we are done. □

Consider any function f and any pure-strategy equilibrium (σ^*, η^*) of the f-context disclosure game. Let $d_1, \ldots, d_N$ index the disclosures that have positive probability under σ^* (i.e., all $d \in \mathcal{H}_n$ such that $\sigma^*(x_n) = d$ for some x_n). For each such disclosure d_i,

$$Z_{\eta^*}(d_i) = \frac{1}{|\{x : \sigma^*(x) = d_i\}|} \sum_{x:\sigma^*(x)=d_i} f(x)$$

is the evaluator's posterior expectation upon observing disclosure d_i. Given the evaluator's payoff function, the optimal action for the evaluator is precisely $Z_{\eta^*}(d_i)$. Let

$$d^* = (H^*, (\mathcal{X}_i^*)_{i \in H^*}) := \arg\max_{1 \leq i \leq N} \phi(Z_{\eta^*}(d_i)) \tag{Q.1}$$

be the disclosure that yields the highest payoff to the Sender. Then it must be that $\sigma^*(x_n) = d^*$ for every covariate vector x_n for which disclosure d^* is feasible. Otherwise d^* would be a profitable deviation. Hence the evaluator's posterior expectation in this equilibrium is the same as it would have been given disclosure of d^* in our main model. So

$$\phi(Z_{\eta^*}(d^*)) \leq \max_{x_n \in \mathcal{X}_n} v(f, x_n).$$

Since the payoff received by an agent with any other covariate vector cannot exceed $\phi(Z_{\eta^*}(d^*))$ (by (Q.1)), we have the desired result.

Q.3 Result for Mixed Strategy Equilibria

In this part we restrict attention to equilibria (σ^*, η^*) with the property that $\arg\max_{\hat{y} \in A_{(\sigma^*,\eta^*)}} \phi(\hat{y})$ is unique on the set $A_{(\sigma^*,\eta^*)}$ of posterior expectations with positive probability in this equilibrium. Call these equilibria *generic*. (A sufficient condition for all equilibria to be generic is if u is strictly monotone.)

For each n and f, let $v^D(f, x_n)$ denote the highest payoff that an agent with covariate vector x_n receives in any generic equilibrium (potentially mixed) of the f-context disclosure game. Further define

$$v_f^D(n) = \max_{x_n} v^D(f, x_n)$$

and

$$V^{\mathcal{D}}(n) = \mathbb{E}[v_f^D(n)]$$

where the expectation is with respect to the realization of f.

Proposition Q.2. *Suppose Assumption 1 holds and $\phi(.)$ is twice continuously differentiable. Then $\lim_{n\to\infty} V^{\mathcal{D}}(n) = 0$.*

Proof. Fix n, f, and a context equilibrium (σ^*, η^*) of the f-context disclosure game. Let $\mathcal{Z}^* \subseteq [-\overline{y}, \overline{y}]$ be the compact set of all equilibrium posterior expectations that are realized with positive probability in this equilibrium. Further, denote

$$Z_{(1)}^* = \arg\max_{z \in \mathcal{Z}^*} \phi(z)$$

to be the most-preferred achievable posterior expectation, which is unique by assumption of genericity of the equilibrium.

Since $Z_{(1)}^*$ is the best attainable posterior expectation, an agent achieves $Z_{(1)}^*$ in equilibrium if and only if it is feasible. (Otherwise, the agent can profitably deviate to the feasible disclosure that induces this posterior expectation.)

Let $\mathcal{X}^* \subseteq \mathcal{X}_n$ denote the set of agents who have a feasible disclosure that achieves $Z_{(1)}^*$. Let $\mathcal{D}(\mathcal{X}^*)$ be the set of disclosures that agents in $\mathcal{X}^*$ send with positive probability in equilibrium. By the logic above, $\mathcal{D}(\mathcal{X}^*) \cap \mathcal{D}(\mathcal{X} \setminus \mathcal{X}^*) = \varnothing$. Using the structure of this equilibrium we can write

$$\mathbb{E}[Y] = Z_{(1)}^* p_{\mathcal{X}^*} + (1 - p_{\mathcal{X}^*}) \mathbb{E}[Y | X \notin \mathcal{X}^*] \tag{Q.2}$$

where $p_{\mathcal{X}^*}$ is the ex-ante probability that the agent's covariate vector belongs to $\mathcal{X}^*$, and $\mathbb{E}[Y|X \notin \mathcal{X}^*]$ is the expectation of the agent's type given that his covariate vector does not belong to $\mathcal{X}^*$. Here we utilize the fact that the evaluator's posterior expectation is constant at $Z^*_{(1)}$ across all agents with covariate vectors in $\mathcal{X}^*$.[27]

Now, consider the following alternative "strategy" σ_0, which relaxes the feasibility constraint: For any $x \in \mathcal{X} \setminus \mathcal{X}^*$ let $\sigma_0(\mathbf{x}) \equiv \sigma^*(\mathbf{x})$, i.e., the disclosures are the same as in the original equilibrium. Further choose some arbitrary disclosure $d_0 \in \mathcal{D}(\mathcal{X}^*)$ and let $\sigma_0(\mathbf{x}) = d_0$ for all $x \in \mathcal{X}^*$. The Receiver's posterior expectation following observation of disclosure d_0 is

$$Z_0 = \frac{\sum_{x \in \mathcal{X}^*} Y_x}{|\mathcal{X}^*|}$$

and, analogous to (Q.2), we can write

$$\mathbb{E}[Y] = Z_0 p_{\mathcal{X}^*} + (1 - p_{\mathcal{X}^*})\mathbb{E}[Y|X \notin \mathcal{X}^*] \tag{Q.3}$$

Combining equations (Q.2) and (Q.3) we conclude:

$$Z^*_{(1)} = \frac{\sum_{x \in \mathcal{X}^*} Y_x}{|\mathcal{X}^*|}$$

which almost surely converges to $\mathbb{E}[Y]$ so long as $|\mathcal{X}^*| \xrightarrow{n\to\infty} \infty$. Since the Y_x's are uniformly bounded, this also implies $\mathbb{E}[Z^*_{(1)}] \to \mathbb{E}[Y]$, as desired. We now demonstrate that indeed $|\mathcal{X}^*| \xrightarrow{n\to\infty} \infty$.

For any disclosure d denote by $C_d \subseteq \mathcal{X}_n$ the set of all covariate vectors $\mathbf{x}$ given which d is feasible. Since $Z^*_{(1)}$ is achieved by all agents for whom $Z^*_{(1)}$ is feasible, it must be that for every disclosure $d \in \mathcal{D}(\mathcal{X}^*)$ we have $C_d \subseteq \mathcal{X}^*$. Then for any $d \in \mathcal{D}(\mathcal{X}^*)$,

$$|\mathcal{X}^*| \geq |C_d| \xrightarrow{n\to\infty} \infty.$$

where the limit follows by assumption that $\alpha_h < 1$. This completes the proof. □

Q.4 Proof of Proposition 2

Throughout the proof we assume $\phi(x) \equiv x$ and $s = 0$. (With general $\phi(x)$, the proof would have an extra step—similar to the proof of Theorem 1, we would use Lipschitz continuity of ϕ to upper bound the expected payoff gain using the expected evaluation gain.)

[27] In general this does not have to be the case. We rule this out in the definition of the equilibrium.

As before, enumerate feasible disclosures by k and denote the corresponding posteriors (as random variables) as $Z_k^n := \rho_f(H_k)$. To upper bound the value of context, we apply a result from Arnold and Groeneveld (1979):

$$\left|\mathbb{E}\left[\max_{k\in\{1,\dots,K_n\}} Z_k^n - \mathbb{E}\left[\frac{\sum_{i=1}^{K_n} Z_i^n}{K_n}\right]\right]\right| \leq \sqrt{\left(1-\frac{1}{K_n}\right)\sum_{i=1}^{K_n} Var(Z_i^n) + \frac{1}{K_n}\sum_{i=1}^{K_n}\left(\sqrt{K_n}\left(\mathbb{E}[Z_i^n] - \frac{\sum_{i=1}^{K_n}\mathbb{E}[Z_i^n]}{K_n}\right)\right)^2} \tag{Q.4}$$

By Assumption 3, inequality Q.4 simplifies to

$$\left|\mathbb{E}\left[\max_{k\in\{1,\dots,K_n\}} Z_k^n\right] - \mu\right| \leq \sqrt{\left(1-\frac{1}{K_n}\right)\sum_{i=1}^{K_n} Var(Z_i^n)}$$

Finally, Assumption 4 implies that $Var(Z_k^n) = o(\frac{1}{K_n})$ uniformly for every disclosure k. Hence

$$\left|\mathbb{E}\left[\max_{k\in\{1,\dots,K_n\}} Z_k^n\right] - \mu\right| \leq \sqrt{\left(1-\frac{1}{K_n}\right) K_n o(K_n^{-1})}$$

which yields the desired result after taking a limit in n. The argument for the lower bound follows the same line of reasoning and is thus omitted.

Q.5 Proof of Proposition 3

Define

$$u_{\max}(f) = \max_{x\in\mathcal{X}_n} \max_{H\subseteq S\cup\mathcal{N}, |H|\leq\alpha_h n}$$

so that $U_n^{\max} = \mathbb{E}_f[u_{\max}(f)]$. Then

$$\begin{aligned}
u_{\max}(f) &= \max_{x\in\mathcal{X}_n} \max_{H_s\subseteq\mathcal{S}, H_n\subseteq\mathcal{N}, |H_s\cup H_n|\leq\alpha_h n} U_x^f(H_s\cup H_n) \\
&\leq \max_{x\in\mathcal{X}_n} \max_{H_s\subseteq\mathcal{S}, H_n\subseteq\mathcal{N}, |H_n|\leq\alpha_h n} U_x^f(H_s\cup H_n) \\
&\leq \max_{x\in\mathcal{X}_n} \max_{H_s\subseteq\mathcal{S}, H\subseteq\mathcal{N}, |H_n|\leq\alpha_h n} U_x^f(H_s\cup H_n) \\
&\leq \max_{x\in\mathcal{X}_n} \left(\max_{H_s\subseteq\mathcal{S}, H_n\subseteq\mathcal{N}, |H_n|\leq\alpha_h n} U_x^f(H_s\cup H_n) - U_x^f(S)\right) + \max_{x\in\mathcal{X}_n} U_x^f(S) \\
&\leq \max_{x\in\mathcal{X}_n} \max_{H_s\subseteq\mathcal{S}, H_n\subseteq\mathcal{N}, |H_n|\leq\alpha_h n} \left|U_x^f(H_s\cup H_n) - U_x^f(S)\right| + \max_{x\in\mathcal{X}_n} U_x^f(S) \\
&\leq B\cdot \max_{x\in\mathcal{X}_n} \max_{H_s\subseteq\mathcal{S}, H_n\subseteq\mathcal{N}, |H_n|\leq\alpha_h n} \left|\hat{f}_{H_s\cup H_n}(x) - \hat{f}_S(x)\right| + \max_{x\in\mathcal{X}_n} U_x^f(S)
\end{aligned}$$

for some positive constant B, where the final inequality follows from Lipschitz continuity of ϕ. Since the absolute value is a quasi-convex function, the largest absolute difference $\left|\hat{f}_{H_s \cup H_n}(x) - \hat{f}_S(x)\right|$ is attained when $H_s \equiv S$ for some covariate vector x. Thus

$$B \cdot \max_{x \in \mathcal{X}_n} \max_{H_s \subseteq \mathcal{S}, H_n \subseteq \mathcal{N}, |H_n| \leq \alpha_h n} \left|\hat{f}_{H_s \cup H_n}(x) - \hat{f}_S(x)\right| + \max_{x \in \mathcal{X}_n} U_x^f(S) \tag{Q.5}$$

$$= B \cdot \max_{x \in \mathcal{X}_n} \max_{H_n \subseteq \mathcal{N}, |H_n| \leq \alpha_h n} \left|\hat{f}_{S \cup H_n}(x) - \hat{f}_S(x)\right| + \max_{x \in \mathcal{X}_n} U_x^f(S) \tag{Q.6}$$

Consider the first term in (Q.6). Theorem 1 establishes that the expected maximum value of context vanishes as n grows large. As a corollary of this result we have:

$$\mathbb{E}\left[\max_{x \in \mathcal{X}_n} \max_{H_n \subseteq \mathcal{N}, |H_n| \leq \alpha_h n} \left|\hat{f}_{S \cup H_n}(x) - \hat{f}_S(x)\right|\right] \to 0 \tag{Q.7}$$

Now consider the second term in (Q.6). Disclosing a vector of standard covariates x_S induces a posterior $\hat{f}_\varnothing(x_S) \xrightarrow{n \to \infty} \mathbb{E}[\hat{f}_\varnothing(x_S)]$. Since S is a finite set, by the continuous mapping theorem we can combine this result with (Q.5) and (Q.7) to get

$$\lim_{n \to \infty} \mathbb{E}[u_{\max}(f)] \leq \max_{x_S \in \{0,1\}^s} \phi(\mathbb{E}[\hat{f}_\varnothing(x_S)])$$

Now we prove a lower bound. Since $H_s = S$ is a feasible disclosure, we get:

$$\max_{x \in \mathcal{X}_n} U_x^f(S) \leq u_{\max}(f)$$

Taking an expectation with respect to f, combining this inequality with the upper bound above, and then taking the limit as n grows, we obtain

$$\max_{x_S \in \{0,1\}^s} \phi(\mathbb{E}[\hat{f}_\varnothing(x_S)]) \leq \lim_{n \to \infty} \mathbb{E}[u_{\max}(f)] \leq \max_{x_S \in \{0,1\}^s} \phi(\mathbb{E}[\hat{f}_\varnothing(x_S)])$$

Or, simply

$$\lim_{n \to \infty} \mathbb{E}[u_{\max}(f)] = \max_{x_S \in \{0,1\}^s} \phi(\mathbb{E}[\hat{f}_\varnothing(x_S)])$$

as desired.