The Value of Context: Human versus Black Box Evaluators

Andrei Iakovlev and Annie Liang*

February 17, 2024

Abstract

Evaluations once solely within the domain of human experts (e.g., medical diagnosis by doctors) can now also be carried out by machine learning algorithms. This raises a new conceptual question: what is the difference between being evaluated by humans and algorithms, and when should an individual prefer one form of evaluation over the other? We propose a theoretical framework that formalizes one key distinction between the two forms of evaluation: Machine learning algorithms are standardized, fixing a common set of covariates by which to assess all individuals, while human evaluators customize which covariates are acquired to each individual. Our framework defines and analyzes the advantage of this customization—the *value of context*—in environments with very high-dimensional data. We show that unless the agent has precise knowledge about the joint distribution of covariates, the value of more covariates typically exceeds the value of context.

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*

Predictions about people are increasingly automated using black-box algorithms. How should individuals compare evaluation by algorithms (e.g., medical diagnosis by a machine

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

learning algorithm) with more traditional evaluation by human experts (e.g., medical diagnosis by a doctor)?

One important distinction is that black-box algorithms are standardized, fixing a common set of inputs by which to assess all individuals. Unless the inputs to the black box are exhaustive, additional information can (in some cases) substantially modify the interpretation of those inputs that have been acquired. For example, the context that a patient is currently fasting may change the interpretations of "dizziness" and "electrolyte imbalance," and the context that a job applicant is an environmental activist may change how a prior history of arrest is perceived. If these auxiliary characteristics are not specified as inputs in the algorithm, the individual cannot supply them.

In contrast, individuals can often explain their unusual circumstances or characteristics to a human evaluator through conversation. Thus, even if the human evaluator considers fewer inputs than a black box algorithm does, these inputs may be better adapted to the individual being evaluated. Longoni et al. (2019) report that the perception that humans are better able to take into account an individual's unique situation contributes significantly to patient resistance to AI in healthcare. Our objective is to understand when, and to what extent, this difference matters.

Our contribution in this paper is twofold. First, we propose a theoretical framework that formalizes this distinction between human and black box evaluation. Second, we explain assumptions under which it will turn out that the the agent should prefer one form of evaluation over the other. We see our paper as a complement to a growing empirical literature that compares human versus black box evaluation. Here our goal is to conceptualize the difference between human and black box evaluators, and to clarify properties of the informational environment that are important for choosing between the two.

In our model, an agent is described by a binary covariate vector and a real-valued type (e.g., the severity of the agent's medical condition). The type can be written as a function of the covariates, which we henceforth call the *type function*. Covariates are separated into standard covariates (e.g., medical history, lab tests, imaging scans) and nonstandard covariates (e.g., religious information, genetic data, wearable device data, and financial data). Since in principle there is no limit on the number of nonstandard covariates that can describe a person, our results consider asymptotics as the number of nonstandard covariates grows to infinity.

We suppose that the agent may know how the standard covariates are correlated with the type, but cannot distinguish between the predictive roles of the nonstandard covariates. Formally, the agent has a belief over the type function, and we impose an exchangeability assumption that says that the agent's prior over these functions is unchanged by permuting the labels and values of the nonstandard covariates. If we interpret the covariates as signals about the agent's type, then uncertainty about the type function corresponds to uncertainty about the signal structure (à la model uncertainty, e.g., Acemoglu et al. (2015) and Morris and Yildiz (2019)).

The agent's payoff is determined by his true type and an evaluation, which may be made either by a human evaluator or a black-box evaluator. In either case the evaluation is a conditional expectation of the agent's type given the agent's standard covariates and some fraction of the agent's nonstandard covariates. But the sets of nonstandard covariates that are observed by the black box evaluator and the human evaluator differ in two ways.

First, the black box evaluator observes a larger fraction of the nonstandard covariates than the human evaluator does. Second, the nonstandard covariates observed by the black box evaluator are a pre-specified set of algorithmic inputs, which are fixed across individuals. For example, a designer of a medical algorithm may specify a set of inputs including (among others) blood type, BMI, and smoking status, and train a black box algorithm to learn the mapping from those inputs into the diagnosis. We view the human evaluator as instead uncovering nonstandard covariates during a conversation, where the specific path of questioning may vary across agents. Thus the human evaluator may end up learning about one individual's sleep schedule and another's financial situation, where the final set of observed covariates is a function of the agent's covariate vector.

Rather than modeling these conversations directly, we consider an upper bound on the agent's payoff under human evaluation, where the covariates that the human observes are the ones that maximize the agent's payoffs (subject to the human's capacity constraint). We say that the agent prefers the black box if the agent's expected payoffs are higher under black box evaluation even compared to these *best-case* conversations with the human.

This comparison essentially reduces to the question of whether the agent prefers an evaluator who observes a larger fraction of (non-targeted) nonstandard covariates about the agent, or an evaluator who observes a smaller but targeted fraction of nonstandard covariates. Towards this comparison, we first introduce a benchmark, which is the expected payoff that the agent would receive if interacting with an evaluator who observes no nonstandard covariates. We define the *value of context* to be the improvement in the agent's payoffs under best-case human evaluation, relative to this benchmark. The value of context thus quantifies the extent to which the agent's payoffs can possibly be improved when the evaluator observes nonstandard covariates suited to that agent.

Our first main result says that under our symmetry assumption on the agent's prior, the expected value of context vanishes to zero as the number of covariates grows large. Thus even though there may be realizations of the type function 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 in some application, it must be that our symmetry assumption does not hold, i.e., 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 expected 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 use this result to compare the agent's expected payoff under human and black box evaluation. We show that when the agent prefers a more accurate evaluation—formally, when the agent's payoff is convex in the evaluation—the agent should prefer an algorithmic evaluator with access to more covariates over a human evaluator to whom the agent can provide context. And when the agent's payoff is concave in the evaluation, the conclusion is reversed.

We subsequently strengthen our main results in two 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. Thus, the expected value of context is eventually small for everyone in the population. 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).

We conclude by examining the role of the symmetry assumption on the agent's prior, and the extent to which our results depend upon it. First, we study two variations of our main model, in which the symmetry assumption is relaxed: In the first, we suppose that there is a "low-dimensional" set of covariates that relevant for predicting the agent's type; in the second, we suppose that the agent knows ex-ante the predictive role of certain nonstandard covariates. In both of these settings, our main results extend partially but can also fail: For example, if the set of relevant covariates is sufficiently small that they can be fully disclosed to the evaluator, then the expected value of context typically will not vanish. Finally we show that precise symmetry is not important for our result—it is enough if the informativeness of each individual set of covariates vanishes as the total number of covariates grows large. Together with our main results, these extensions clarify different categories of informational assumptions under which the expected value of context does or does not turn out to be high.

Our model is not meant to be a complete description of the differences between human and black box evaluation. For example, we have abstracted away from human and algorithmic bias (Kleinberg et al., 2017; Gillis et al., 2021), factors such as empathy, and the possibility that the human evaluator has access to information that is not available to the algorithm (e.g., for privacy protection 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)).[1] We leave extensions of our model that include these other differences to future work.

1.1 Related Literature

A large literature compares 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). These results have led some to predict or call for the replacement of human evaluation with algorithmic evaluation (Obermeyer and Emanuel, 2016). But human evaluation and/or human oversight of algorithmic predictions remains the norm, in part because of user distrust of algorithmic predictions (Jussupow and Heinzl, 2020; Bastani et al., 2022; Lai et al., 2023).

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).[2] 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,

[1] 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 black box evaluation, which are difficult to resolve by training the algorithm differently.

[2] 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)).

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.

Our formal results are related to the literature on asymptotic learning and agreement across Bayesian agents (Blackwell and Dubins, 1962; Cripps et al., 2008; Acemoglu et al., 2015). Specifically, one can view our main result as bounding (in expectation) the differences in beliefs across Bayesian agents who are given different information. But the asymptotics that we look at are of a different nature from those studied previously. Among other distinctions, we consider asymptotics with respect to a sequence of varying information structures, rather than studying asymptotic beliefs as the total amount of information accumulates.

The agent in our framework has model uncertainty (Acemoglu et al., 2015; Morris and Yildiz, 2019), and the central Assumption 1 constrains the agent's model uncertainty to take a particular (and new) form motivated by the applications we have in mind. The presence of model uncertainty distinguishes our problem from the related and very interesting work of 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. 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.[3]

Finally, our work builds on the literature on persuasion via strategic information disclosure (e.g., Glazer and Rubinstein (2004), Kamenica and Gentzkow (2011)). The model that we study—in which the sender has private information about his type 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).[4] Different from this literature, our sender has uncertainty about how his reports are interpreted, and our focus is not on examining which incentive-compatible reporting strategy is optimal. (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 4.2 that our results extend in a disclosure game.) Our focus is instead on asymptotic limits of belief manipulability as the number of components in the type vector grows large, which is special to our motivation.

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

[3]This distinction is important: If, similar to Di Tillio et al. (2021), we modeled covariates as IID signals about the agent's type, then the expected value of context would not vanish as the number of covariates grew large. See Section 3.2 for more detail.

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

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]

2 Model

2.1 Setup

Agents are each described by a binary covariate vector $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ and a type $y \in [-\overline{y}, \overline{y}]$ (where $0 \leq \overline{y} < \infty$), which are structurally related by the function

$$y = f(x_1, \ldots, x_n).$$

We refer to f henceforth as the *type function.* The distribution over covariate vectors is uniform in the population.[7]

We refer to the covariates indexed to $\mathcal{S} = \{1, \ldots, s\}$ as *standard* covariates and the covariates indexed to $\mathcal{N} = \{s+1, \ldots, n\}$ as *nonstandard* covariates. Since in principle there is no limit on the number of covariates that can describe a person, we view the case of infinite covariates as the relevant one, and our results focus on asymptotics as n grows large.[8]

Example 1 (Job Interview). Standard covariates describing a job applicant may include their work history, education level, college GPA, and the coding languages they know. Nonstandard covariates may include their social media activity (e.g., number of followers, posts,

[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 decision.

[7]All of our results extend for arbitrary finite-valued covariates.

[8]We consider a generalization of our model in Section 5.1 in which a "low-dimensional" subset of these covariates are sufficient for predicting the agent's type.

likes), wearable device data (e.g., sleep patterns, physical activity level), and hobbies (e.g., whether they are active readers, whether they enjoy extreme sports).

Example 2 (Medical Prediction). Standard covariates describing a patient may include symptoms, prior diagnoses, family medical history, lab tests and imaging results. Nonstandard covariates may include the patient's religious practices, genetic data, wearable device data, and financial data.[9]

An *evaluation* of the agent, $\hat{y} \in [-\overline{y}, \overline{y}]$, is described in the following section. The agent has a Lipschitz continuous utility function $u : [-\overline{y}, \overline{y}]^2 \to \mathbb{R}$, which maps the evaluation $\hat{y}$ and the agent's true type y into a payoff.

Example 3 (Higher Evaluations are Better). The agent's payoff is

$$u(\hat{y}, y) = \phi(\hat{y})$$

for some increasing ϕ. This corresponds, for example, to an agent receiving a desired outcome (e.g., a loan or a promotion) with probability increasing in the evaluation.

Example 4 (More Accurate Evaluations are Better). The agent's payoff is

$$u(\hat{y}, y) = -(\hat{y} - y)^2.$$

This corresponds to harms that are decreasing in the accuracy of the evaluation, e.g., medical prediction problems where more accurate evaluations are desired.

2.2 Evaluation of the agent

There are two evaluators: a black box evaluator, henceforth Black Box (it), and a human evaluator, henceforth Human (she). Both evaluators form an evaluation as an expectation of the agent's type y given observed covariates, so we will introduce notation for these conditional expectations. For any covariate vector $\mathbf{x}_n = (x_1, \ldots, x_n)$ and subset of nonstandard covariates $A \subseteq \mathcal{N}$, let

$$C_A(\mathbf{x}_n) = \{\tilde{x} \in \{0,1\}^n : \tilde{x}_i = x_i \quad \forall i \in \mathcal{S} \cup A\} \tag{1}$$

be the set of all covariate vectors that agree with $\mathbf{x}_n$ on the covariates with indices in $\mathcal{S} \cup A$. Further define

$$\hat{y}^f_{\mathbf{x}_n}(A) = \frac{1}{|C_A(\mathbf{x}_n)|} \sum_{x \in C_A(\mathbf{x}_n)} f(x) \tag{2}$$

[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.

to be the conditional expectation of the agent's type given their standard covariates and their nonstandard covariates with indices in A. We use

$$U^f_{\mathbf{x}_n}(A) = u\left(\hat{y}^f_{\mathbf{x}_n}(A), y\right)$$

to denote the agent's payoff given this evaluation.

Both the human and black box evaluation take the form (2), but the observed sets of nonstandard covariates A are different across the evaluators. Black Box observes the nonstandard covariates in the set $B = \{s+1, \ldots, s+b_n\}$ where $b_n = \lfloor \alpha_b \cdot n \rfloor$.[10] Importantly, this set is held fixed across agents. So an individual with covariate vector $\mathbf{x}_n$ receives the evaluation $\hat{y}^f_{\mathbf{x}_n}(B)$ and payoff $U^f_{\mathbf{x}_n}(B)$ when evaluated by the Black Box.[11]

Human differs from Black Box in two ways. First, Human has a capacity of $h_n = \lfloor \alpha_h \cdot n \rfloor$ nonstandard covariates per agent, where $\alpha_h < \alpha_b$ (i.e., Human cannot process as many inputs as Black Box). Second, Human does not pre-specify which nonstandard covariates to observe, but rather learns these through conversation, and thus potentially observes different nonstandard covariates for each agent. For example, a doctor (evaluator) may pose different questions to different patients (agents) depending on their answers to previous questions. Or a job candidate (agent) might choose to disclose to an interviewer (evaluator) certain nonstandard covariates that put him in a good light.

Rather than modeling the complex process of a conversation, we study the quantity

$$\max_{H \subseteq \mathcal{N}, |H| \leq \alpha_h \cdot n} U^f_{\mathbf{x}_n}(H) \tag{3}$$

which is the agent's payoff when the posterior expectation about his type is based on those $\alpha_h \cdot n$ or fewer covariates that are best for him.

We can interpret this quantity as an upper bound for the agent's payoffs under certain assumptions. First, if the evaluator selects which covariates to observe, then (3) is an upper bound on the agent's possible payoffs across all possible evaluator selection rules. Second, if covariates are disclosed by the agent, but the evaluator updates to the disclosed covariates as if they had been chosen exogenously, then again (3) represents an upper bound on the agent's possible payoffs.[12]

[10] One can instead assume that these nonstandard covariates are selected uniformly at random. This will not affect the results of this paper.

[11] It is not important for our results that B is common across individuals; what we require is that any randomness in B is independent of the agent's covariates and type. For example, if the set B were drawn uniformly at random for each agent, our results would hold.

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

If however the covariates are disclosed by the agent in a disclosure game, and the evaluator accounts for the strategic nature of this disclosure, then whether (3) represents an upper bound will depend on what we assume that the agent knows at the time of disclosure.[13] We show in Section 4.2 that if the agent knows his entire covariate vector, then (3) need not upper bound every agent's payoffs. Nevertheless, we present a different quantity that does upper bound the maximum payoff that any agent can obtain in this disclosure game, and show that our main results extend when we replace (3) with this quantity. To streamline the exposition we focus on the prior two interpretations (in which the human evaluator either selects the covariates herself or updates to the agent's disclosures naively), and postpone the discussion of disclosure games to Section 4.2.

2.3 Value of context

A key input towards understanding the comparison between Human and Black Box is quantifying the extent to which individualized context improves the agent's payoffs.

Definition 1 (Value of Context)**.** *The* value of context *for an agent with covariate vector* $\mathbf{x}_n$ *and type* $y = f(\mathbf{x}_n)$ *is*

$$v(f, \mathbf{x}_n) = \max_{H \subseteq \mathcal{N}, |H| \leq \alpha_h n} U^f_{\mathbf{x}_n}(H) - U^f_{\mathbf{x}_n}(\varnothing)$$

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.*

In general, the value of context depends on the type function f as well as on the agent's own covariate vector $\mathbf{x}_n$.[14]

Example 5 (High Value of Context). Let $u(\hat{y}, y) = \hat{y}$, i.e., the agent's payoff is the evaluation. Suppose x_1 is a standard covariate (observed no matter what), while $x_2, \ldots, x_{100}$ are nonstandard covariates. The type y is related to these covariates via the type function

$$y = f(x_1, \ldots, x_{100}) = \begin{cases} c & \text{if } x_1 = x_2 \\ -c & \text{if } x_1 \neq x_2 \end{cases}$$

For an agent who can reveal (up to) one covariate and whose covariate vector is $(1, 1, \ldots, 1)$, the value of context is c, since revealing $x_2 = 1$ moves the expectation of his type from 0

[13] This is roughly because the agent can potentially "sneak in" information about the other covariates via the covariates that are revealed.

[14] The value of context given a specific function f is spiritually related to the communication complexity of f (Kushilevitz and Nisan, 1996).

to c. This example corresponds to settings in which some nonstandard covariate substantially moderates the interpretation of a standard covariate. For such type functions f, it is important for the evaluator to observe the right nonstandard covariates, and so the value of context can be large.

Example 6 (Low Value of Context). Suppose the type function in the previous example is instead $y = f(x_1) = x_1$ (leaving all other details of the example unchanged). Then the value of context is 0 for every agent. In this example, nonstandard covariates are irrelevant for predicting the type, so there is no value to the evaluator discovering the "right" covariates.

We are interested in settings where the agent does not know the type function, and hence cannot compute the value of context. We give the agent uncertainty about f and characterize the agent's expected value of context and expected payoffs, integrating over the agent's belief about f.[15]

We do this for two reasons. First, in many applications it is not realistic to suppose that the agent knows f. For example, a patient who anticipates that a diagnosis will be based on an image scan of his kidney may recognize that there are properties of the image that are indicative of whether he has the condition or not, but likely does not know what the relevant properties are, or how they determine the diagnosis.[16]

Second, it turns out that the case with uncertainty about f yields a more elegant analysis. Although the value of context can be computed for specific f, it typically depends on details of that function, as well as on the agent's own covariate vector. In contrast, under a condition on the prior belief (defined in the following section) it is possible to draw strong detail-free conclusions.

2.4 Model Uncertainty

Our leading assumption is that the agent may know how standard covariates impact the type, but has no ex-ante knowledge about the different roles of the nonstandard covariates. Formally, we impose the following.

Definition 2 (Finite Exchangeability)**.** *The sequence of random variables $W_1, W_2, \ldots, W_n$ is* finitely exchangeable *if for every permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, the sequences $(W_1, \ldots, W_n)$ and $(W_{\pi(1)}, \ldots, W_{\pi(n)})$ have the same joint distribution.*

[15] If we interpret the covariates in our model as signals about the type, then the function relating covariates to type corresponds to the signal structure.

[16] In the case of a job interview, the function f may reflect particular subjective preferences of the firm, which are initially unknown to the agent.

Definition 3 (Infinite Exchangeability). *The sequence of random variables* $W_1, W_2, \ldots$ *is infinitely exchangeable if for every* $n \in \mathbb{N}$ *and set of indices* $i_1, \ldots, i_n$*, the sequence* $(W_{i_1}, \ldots, W_{i_n})$ *is finitely exchangeable.*

In what follows, for each covariate vector $\mathbf{x}$ we define the random variable $Y_{\mathbf{x}} = f(\mathbf{x})$ (where the randomness is through the type function f).

Assumption 1. *Fix any realization of the standard covariates* $\mathbf{x}_{\mathcal{S}} = (x_1, \ldots, x_s) \in \{0,1\}^s$*. There is an infinitely exchangeable sequence of* $[-\overline{y}, \overline{y}]$*-valued random variables* $(\widetilde{Y}_1, \widetilde{Y}_2, \ldots)$ *such that for every* $n \in \mathbb{N}$*, the sequence*

$$(Y_{\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}}} : \mathbf{x}_{-\mathcal{S}} \in \{0,1\}^{n-s})$$

has the same distribution as $(\widetilde{Y}_1, \ldots, \widetilde{Y}_{2^{n-s}})$*.*

The sequence $(Y_{\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}}} : \mathbf{x}_{-\mathcal{S}} \in \{0,1\}^{n-s})$ includes all types associated with the covariate vectors that "complete" $\mathbf{x}_{\mathcal{S}}$ by filling in values for the nonstandard covariates. Assumption 1 says that the joint distribution of these types is ex-ante invariant to permutations of the covariate vectors within the set $\{(\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}}) : \mathbf{x}_{-\mathcal{S}} \in \{0,1\}^{n-s}\}$. An agent whose prior is given by this joint distribution is thus agnostic about how the labels and values of the nonstandard covariates impact the type.

Besides imposing ex-ante symmetry of the nonstandard covariates, the main content of the assumption is that the unconditional distribution of y in the population of agents is constant across n.[17] If we interpret each covariate as a bit of information, then under Assumption 1, the number of covariates n can be interpreted as moderating the richness of the informational environment and the potential complexity of the mapping f, but not as a measure of the quantity of information.[18] In other words, as n grows large, the agent has a more extensive set of words to describe a fixed unknown y.

Some simple examples of priors satisfying this assumption are given below. These examples clarify that symmetry of the nonstandard covariates is imposed only on the agent's prior belief, and not on the realized f.[19]

Example 7. Let $y \in \{0,1\}$, in which case the space of possible functions $f : \mathcal{X} \to \mathcal{Y}$ 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.

[17] This contrasts, for example, with a model in which $x_1, x_2, \ldots$ were drawn i.i.d. from a type-dependent distribution F_y. The total quantity of information about y would then be increasing in the number of covariates. In our model it is not.

[18] As we show in Appendix A.1, it is possible to extend our framework so that the type is random conditional on the covariate vector $\mathbf{x}_n$; what is crucial is that the size of residual uncertainty does not depend on n.

[19] That is, we do not restrict the agent's prior belief to have support on functions f where $f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)})$ for permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Example 8. Suppose there is a distribution F on $[-\overline{y}, \overline{y}]$ such that for each n,

$$\left(f(\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}}) : \mathbf{x}_{-\mathcal{S}} \in \{0,1\}^{n-s}\right) \sim_{i.i.d.} F.$$

Then Assumption 1 is satisfied.

Assumption 1 describes settings in which many different structural relationships between the covariates and the type are plausible (including both ones where the value of context will turn out to be high and low), but ex-ante those relationships are not known. It thus rules out prior knowledge about asymmetries across covariates or covariate values, such as in the following examples.

Example 9 (Only One Covariate is Relevant). The type is equal to the value of nonstandard covariate x_I, where the index I is drawn uniformly at random from $\mathcal{N}$.

Example 10 (Higher Values are Better). The value of $f(\mathbf{x}_n)$ is (independently) drawn from a uniform distribution on $[1, 2]$ if $x_{s+1} = 1$, and (independently) drawn from a uniform distribution on $[0, 1]$ if $x_{s+1} = 0$.

We view Assumption 1 as a useful conceptual benchmark, and later explore how far our main results generalize under various relaxations of this assumption (Section 5). While our assumption of no prior knowledge about the role of nonstandard covariates is strong, by definition the nonstandard covariates are precisely the covariates for which there is little historical data. For example, while it may be well understood that a higher GPA correlates with higher on-the-job ability, a large number of social media followers could potentially be a positive or negative signal. Indeed, if it were understood that some nonstandard covariate was very predictive—for example, that having many social media followers was a strong indication of on-the-job success—we would expect this nonstandard covariate to become a standard covariate, and thus queried both by the human and black box evaluator.

We now define the expected value of context from the point of view of an agent who knows his covariate vector $\mathbf{x}_n$ but does not know the function f (and hence also does not know his type $y = f(\mathbf{x}_n)$). As we show in Section 4.1, the assumption that the agent knows $\mathbf{x}_n$ is immaterial for the results.

Definition 4 (Expected Value of Context). *For every $n \in \mathbb{Z}_+$ and covariate vector $\mathbf{x}_n \in \{0,1\}^n$, the* expected value of context *is*

$$V(n, \mathbf{x}_n) = \mathbb{E}\left[v(f, \mathbf{x}_n)\right].$$

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

We similarly compare evaluators based on the expected payoff that the agent receives.

Definition 5. *Consider any agent with covariate vector* $\mathbf{x}_n$*. If*

$$\mathbb{E}\left[\max_{H\subseteq\mathcal{N},|H|\le\alpha_h\cdot n} U^f_{\mathbf{x}_n}(H)\right] < \mathbb{E}\left[U^f_{\mathbf{x}_n}(B)\right] \tag{4}$$

then say that the agent prefers the black box evaluator*. And if*

$$\mathbb{E}\left[\min_{H\subseteq\mathcal{N},|H|\le\alpha_h\cdot n} U^f_{\mathbf{x}_n}(H)\right] > \mathbb{E}\left[U^f_{\mathbf{x}_n}(B)\right] \tag{5}$$

then say that the agent prefers the human evaluator*.*

These definitions correspond to a thought experiment in which (for example) a patient has a choice between being seen by a doctor or assessed by an algorithm. If the patient chooses the algorithm, his standard covariates and $\alpha_b \cdot n$ arbitrarily chosen nonstandard covariates will be sent to the algorithm. If the patient chooses the doctor, he will engage in a conversation with the doctor, where his standard covariates and $\alpha_h \cdot n$ selected nonstandard covariates will be revealed. Which should the patient choose?

The first part of Definition 5 compares the agent's expected payoff under black box evaluation with the *best-case* expected payoff under human evaluation, namely when the human evaluator observes those (up to) $\alpha_h \cdot n$ covariates that maximize the agent's payoffs. If the agent's expected payoff is nevertheless higher under black box evaluation even after biasing the agent towards the human in this way, we say that the agent *prefers to be evaluated by the black box.* The second part of the definition compares the agent's expected payoff under black box 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 the agent's expected payoff is lower under black box evaluation even after biasing the agent against the human in this way, then we say that the agent *prefers to be evaluated by the human.*[20]

These are clearly very conservative criteria for what it means to prefer the human or the black box. In practice, we would expect that the set of revealed covariates H to be intermediate to the two cases considered in Definition 5, i.e., that H neither maximizes nor minimizes the agent's payoffs.[21] But if we can conclude either that the agent prefers the black box evaluator or the human evaluator according to Definition 5, then the same conclusion would hold for these more realistic models of H.

[20] In Section 4.2 we further discuss the extent to which these interpretations are valid when the evaluator also updates her beliefs to the selection of covariates.

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

3 Main Results

Section 3.1 characterizes the expected value of context in a simple example. Section 3.2 presents our first main result, which says that the expected value of context vanishes to zero as the number of covariates grows large. Section 3.3 compares human and black box evaluators.

3.1 Example

Suppose there are two covariates x_1 and x_2, both nonstandard. For each covariate vector $\mathbf{x} \in \{0,1\}^2$, define the random variable $Y_{\mathbf{x}} = f(\mathbf{x})$, where the randomness is in the realization of f.

X_1	X_2	$Y_{\mathbf{x}}$
0	0	Y_{00}
0	1	Y_{01}
1	0	Y_{10}
1	1	Y_{11}

Table 1: The four possible covariate vectors and their associated types.

The agent has utility function $u(\hat{y}, y) = \hat{y}$ and covariate vector $(1,1)$. Suppose Human observes up to one nonstandard covariate; then, there are three possibilities for what the evaluator observes. If Human observes $x_1 = 1$, her evaluation is

$$Z_1 \equiv \frac{Y_{10} + Y_{11}}{2}.$$

If Human observes $x_2 = 1$, her evaluation is

$$Z_2 \equiv \frac{Y_{01} + Y_{11}}{2}.$$

And if Human observes no nonstandard covariates, then her evaluation remains the unconditional average

$$Z_\varnothing \equiv \frac{Y_{00} + Y_{01} + Y_{10} + Y_{11}}{4}.$$

So the expected value of context for this agent is

$$\mathbb{E}\left[\max\left\{Z_\varnothing, Z_1, Z_2\right\} - Z_\varnothing\right]. \tag{6}$$

Suppose n grows large with up to $h_n = \lfloor \frac{n}{2} \rfloor$ covariates observed. There are two opposing forces affecting the value of context. First, when n is larger there are more distinct sets of covariates that can be revealed to Human, and hence the max in (6) is taken over a larger number of posterior expectations. This increases the value of context. On the other hand, each Z_k is a sample average, and the number of elements in this sample average also grows in n.[22] By the law of large numbers, each Z_k thus concentrates on its expectation (which is common across k) as n grows large, so the difference between any Z_k and $Z_{k'}$ grows small. What we have to determine is whether the growth rate in the number of subsets of nonstandard covariates (of size $\leq h_n$) is sufficiently large such that the maximum difference in evaluations across these sets is nevertheless asymptotically bounded away from zero. The answer turns out to be no.

3.2 The Expected Value of Context

Our main result says that for every agent, the expected value of context vanishes as n grows large.

Theorem 1. *Suppose Assumption 1 holds. Then for every covariate vector* $\mathbf{x} \in \{0,1\}^\infty$,

$$V(n, \mathbf{x}_n) = O\left(\frac{\sqrt{n}}{2^{(1-\alpha_h)n}}\right) \tag{7}$$

Hence for any $\alpha_h < 1$*, the expected value of context vanishes to zero as n grows large, i.e.,*

$$\lim_{n\to\infty} V(n, \mathbf{x}_n) = 0.$$

Thus although the value of context may be substantial for certain type functions (such as in Example 5), it does not matter on average across these functions when the agent's prior satisfies Assumption 1. This also implies that provision of context does not "typically" matter; that is, the probability that the agent gains substantially from targeted information acquisition is small.

The core of the proof of Theorem 1 is an argument that the extent to which context can change the evaluator's posterior expectation vanishes in the number of covariates. We outline that argument here. For each n, there are $K_n = \sum_{j=0}^{\lfloor \alpha_h n \rfloor} \binom{n-s}{j}$ sets of $\alpha_h n$ (or fewer) nonstandard covariates that can be disclosed. We can enumerate and index these sets to $k = 1, \ldots, K_n$. Each set k induces a posterior expectation Z_k which is a sample average of random variables $Y_x \equiv f(x)$. The expected value of context (for this utility function) is

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

[22] For example, observing $X_1 = 1$ with $n = 2$ gives the evaluator a posterior expectation of $(Y_{10} + Y_{11})/2$, while the same observation gives the evaluator a posterior expectation of $(Y_{100} + Y_{101} + Y_{110} + Y_{111})/2$ if $n = 3$.

where $Z_\varnothing$ is Human's posterior expectation given observation of standard covariates only. Normalizing $E[Z_\varnothing] = 0$, it remains to study properties of the first-order statistic $\max_{1\leq k\leq K_n} Z_k$.

There are two challenges to analyzing this quantity. First, the correlation structure of $Z_1, \ldots, Z_{K_n}$ can be complex: The variables Z_k are neither independent (because the same posterior expectation Y_x can appear as an element in different sample averages $Z_k, Z_{k'}$) 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 $(Z_1, \ldots, Z_{K_n})$ 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 Z_k^{iid}, with the property that

$$\mathbb{E}\left[\max\{Z_1, \ldots, Z_{K_n}\}\right] \leq \mathbb{E}\left[\max\{Z_1^{iid}, \ldots, Z_{K_n}^{iid}\}\right] \tag{8}$$

Applying a result from Chernozhukov et al. (2013), we show that $\max_{1\leq k\leq K_n} Z_k^{iid}$ (properly normalized) converges to $\max_{1\leq k\leq K_n} Z_k^{Normal}$ in distribution, where (due to properties of our problem) $Z_k^{Normal} \sim_{iid} \mathcal{N}\left(0, \frac{1}{2^{n(1-\alpha_h)-s}}\right)$. Having reduced the analysis to studying the expected maximum of i.i.d. Gaussian variables, classic bounds apply to show that this quantity is no more than

$$\frac{1}{2^{n(1-\alpha_h)-s}}\sqrt{\log(K_n)}. \tag{9}$$

This display quantifies the importance of each of the two forces discussed in Section 3.1. First, as n grows larger, the number of posterior expectations $K_n = \sum_{j=0}^{\lfloor \alpha_h n\rfloor}\binom{n-s}{j} \leq 2^{n-s}$ 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, $\frac{1}{2^{n(1-\alpha_h)-s}}$, decreases exponentially in n. What the bound in display (9) 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.

This proof sketch also clarifies the role of Assumption 1. As we show in Section 5.3, the statement of the theorem extends so long as the evaluator's posterior expectation Z_k concentrates on its expectation sufficiently quickly as n grows large. Roughly speaking, this means that the informativeness of any specific set of covariates is decreasing in the total number of covariates. Thus neither the precise symmetry imposed by Assumption 1, nor even the use of Bayesian updating, is critical for a result like Theorem 1 to hold.

On the other hand, the conclusion of Theorem 1 can fail if the agent has substantial prior knowledge about how y is related to the covariates.

Example 11. Let $s = 0$, so that there are no standard covariates. Suppose that for each n,

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

where U is a uniform random variable on $[0, 1]$. This model violates Assumption 1, since it is known that higher realizations of the agent's covariates are good news about the agent's type. The conclusion of Theorem 1 also does not hold: For any n, the evaluator's prior expectation is $\mathbb{E}[f(\mathbf{x}_n)] = 1/4$. But if $\lfloor \alpha \cdot n \rfloor$ covariates are revealed to be 1, the evaluator's posterior expectation is equal to $\frac{1}{4} + \frac{1}{4} \frac{\lfloor \alpha n \rfloor}{n}$. So the expected value of context for an agent with $\mathbf{x}_n = (1, ..., 1)$ is asymptotically bounded away from zero.

In Section 5 we explore several relaxations of Assumption 1: The first of these relaxations supposes that there is a "low-dimensional" set of covariates that predictive of the agent's type, while the remaining covariates are irrelevant. The second relaxation supposes that there is a subset of nonstandard covariates whose effects are known. We formalize these extensions of our main model and examine the extent to which Theorem 1 extends.

3.3 Human versus Black Box

We now turn to the question of when the agent should prefer the human evaluator and when the agent should prefer the black box evaluator.

Assumption 2. *The agent's expected utility can be written as* $\mathbb{E}[\phi(\hat{y})]$ *for some twice continuously differentiable function* ϕ.[23]

Theorem 2. *Suppose Assumptions 1 and 2 hold.*

(a) If ϕ *is strictly convex, then there exists* N *sufficiently large that the agent prefers the black box evaluator for all* $n \geq N$.

(b) If ϕ *is strictly concave, then there exists* N *sufficiently large that the agent prefers the human evaluator for all* $n \geq N$.

Consider first the case of convex ϕ (Part (a)), corresponding to a preference for more accurate evaluations.[24] Such an agent prefers for the evaluation to be based on more infor-

[23] 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).

[24] 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.

mation (advantaging Black Box), but also prefers for the evaluation to be based on more relevant covariates (advantaging Human). We use the rate of convergence demonstrated in Theorem 1 (see (7)) to show that what eventually dominates is how many covariates the evaluators observe, not how they are selected; for an agent who prefers accuracy, this favors the Black Box.

Part (b) of Theorem 2 says that if instead ϕ is concave (corresponding to an agent who prefers inaccurate evaluations) then the agent should eventually prefer the human evaluator. We conclude this section with example decision problems that induce utility functions satisfying the conditions of either part of the theorem.

Example 12. Suppose the agent receives a dollar wage equal to the evaluation, and is risk averse in money. Then his utility function is $u(\hat{y}, y) = \phi(\hat{y})$ for some increasing and concave ϕ, and Part (a) of Theorem 2 says that the agent prefers to be evaluated by the human.

Example 13. Suppose the agent's type is $y \in \{0, 1\}$, and the evaluator chooses an action a based on the observed covariates. The evaluator and agent share the utility function $-\mathbb{E}[(a-y)^2]$. The evaluator's optimal action is $a = \hat{y}$, and the agent's expected payoff given this action is $\mathbb{E}[\hat{y}^2 - \hat{y}]$. So Part (b) of Theorem 2 says that the agent eventually prefers evaluation by the black box evaluator.[25]

4 Extensions

We now show that we are able to strengthen our main results (Theorems 1 and 2) in the following ways. In Section 4.1, we show that not only does the expected value of context vanish for each individual agent, but in fact the expected maximum value of context across agents also vanishes. That is, in expectation the most that context can benefit *any* agent in the population is small. From this, it is immediate that our main results also extend in a generalization of our model in which the agent has uncertainty over his covariate vector. In Section 4.2, 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.

[25] Although the conditions of Theorem 2 are no longer met when y is not binary, we show in Appendix A.6 that the conclusion of Part (b) of Theorem 2 generalizes for arbitrary y given the mean-squared error payoff function described in this example.

4.1 Max value of context across agents

So far we have studied the the expected value of context for a single agent. If we instead ask whether the firm should use human or algorithmic evaluation—for example, whether a hospital should automate diagnoses or rely on doctor evaluations—various other statistics may also be relevant. For example, it may matter whether the value of context is large for any agent in the population (e.g., because a lawsuit regarding algorithmic error may be brought on the basis of harm to any specific individual (Jha, 2020)). We thus study the expected maximum value of context, as defined below.

Definition 6. *For any $n \in \mathbb{Z}_+$, the* expected maximum value of context *is*

$$V^{max}(n) = \mathbb{E}\left[\max_{\mathbf{x}_n \in \{0,1\}^n} v(f, \mathbf{x}_n)\right].$$

The following corollary says that this quantity also vanishes as n grows large.

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

Thus, the expected value of context vanishes uniformly across agents in the population. This result immediately implies that Theorems 1 and 2 extend in any generalization of our model in which the agent has uncertainty not only over f but also over his own covariate vector $\mathbf{x}_n$.

4.2 Strategic Disclosure

So far we've remained agnostic as to whether the agent or evaluator chooses which nonstandard covariates are revealed, 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 which nonstandard covariates are revealed, and the human evaluator updates her beliefs about the agent's type in part based on which covariates are chosen.

For any fixed function f, call the following an *f-context disclosure game*: There are two players, the agent and the evaluator. The function f is common knowledge.[26] 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 $H \subseteq \mathcal{N}$ and values for those covariates. A disclosure $d = (H, (x'_i)_{i\in H})$ is *feasible*

[26] 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, as they are all symmetric from the agent's point of view.

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 $i \in H$.

The agent chooses a *disclosure strategy*, which is a map

$$\sigma : \{0,1\}^n \to \mathcal{D}$$

from covariate vectors to feasible disclosures. The agent then privately observes his covariate vector $\mathbf{x}_n$ and discloses $\sigma(\mathbf{x}_n)$. 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 7. *Let $v^D(f, \mathbf{x}_n)$ denote the highest payoff that an agent with covariate vector $\mathbf{x}_n$ receives in any pure-strategy equilibrium of the f-context disclosure game. The* expected maximum value of context disclosure *is*

$$V^D(n) = \mathbb{E}\left[\max_{\mathbf{x}_n \in \{0,1\}^n} v^D(f, \mathbf{x}_n)\right].$$

We show that the best payoff 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 1, our previous results extend.

5 Relaxing the Symmetry Assumption

As shown in Example 11, our main results can fail if the assumption of symmetric uncertainty over the role of the nonstandard covariate values (Assumption 1) is broken. We now propose three particular variations on Assumption 1 and explore the extent to which our main result extends. In Section 5.1, we suppose that it is known ex-ante that some r_n covariates are relevant, while the remaining $n - r_n$ are not, so that even as n grows to infinity, the effective number of covariates potentially grows more slowly. In Section 5.2 we allow the agent to have prior knowledge about the role of certain nonstandard covariates. Finally, 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.

5.1 Irrelevant covariates

Under Assumption 1, it cannot be known ex-ante that some covariates are irrelevant for predicting the type. The assumption thus rules out settings such as the following.

Example 14. The evaluator is a job interviewer. Although in principle there is an infinite number of covariates that can describe a job candidate, it is understood that not all of them are actually relevant to the job candidate's ability. That is, there is some potentially large (but not exhaustive) set of covariates that contain all of the predictive content about the candidate's ability, and the remaining covariates are either irrelevant for predicting ability, or are predictive only because they correlate with other intrinsically predictive covariates.

If irrelevant covariates cannot be disclosed to the evaluator, then we return to our main model with a smaller n and our previous results extend directly. The more novel case is the one in which it is known that $n - r_n$ covariates are irrelevant, but those covariates can still be disclosed to the evaluator (for example, because it is not commonly understood that they are irrelevant).[27]

To model this, we suppose there is a sequence of sets of *relevant covariates* $(R_1, R_2, \dots)$ such that each R_n includes the standard covariates in $\mathcal{S}$ and is of size $s + r_n$, where $r_n = \lfloor \alpha_r \cdot n \rfloor$ is the (known) number of relevant nonstandard covariates. The irrelevance of the remaining covariates is reflected in the following assumption, which says that, holding fixed the values of the relevant covariates, the values of the irrelevant covariates do not change the type.

Assumption 3 (Irrelevance)**.** *For every $x_{R_n} \in \{0,1\}^{s+r_n}$ and $x_{-R_n}, x'_{-R_n} \in \{0,1\}^{n-s-r_n}$,*

$$f(x_{R_n}, x_{-R_n}) = f(x_{R_n}, x'_{-R_n})$$

We then modify Assumption 1 to impose symmetry only over realizations of the relevant covariates.

Assumption 4. *Fix any realization of the standard covariates $\mathbf{x}_\mathcal{S} \in \{0,1\}^s$. There is an infinitely exchangeable sequence $(\widetilde{Y}_1, \widetilde{Y}_2, \dots)$ such that for every $n \in \mathbb{N}$, the sequence*

$$(Y_{\mathbf{x}_{R_n}, \mathbf{x}_{-R_n}} : (x_i)_{i \in R_n \setminus \mathcal{S}} \in \{0,1\}^{r_n - s})$$

has the same distribution as $(\widetilde{Y}_1, \dots, \widetilde{Y}_{2^{r_n}})$.

Our main model is otherwise unchanged—in particular, we allow the agent to disclose any of the $n - s$ nonstandard covariates, including those which are irrelevant. We show that our previous results extend so long as $\alpha_h < \alpha_r$.

[27] To see the difference, consider the case in which the agent simply wants the evaluator to hold a higher posterior expectation. The irrelevant covariates create noise, and for some realizations of f it may be that disclosing an irrelevant covariate leads to a higher evaluation.

Proposition 2. *Suppose Assumption 1 holds and $\alpha_h < \alpha_r$. Then for every covariate vector $\mathbf{x} \in \{0,1\}^\infty$ the expected value of context vanishes to zero as n grows large.*

The case where $\alpha_r < \alpha_h$ (violating the assumption of the result) corresponds to a setting in which the number of relevant covariates is so small that the agent can disclose all of them. For example, if a job candidate is convinced that only 10 nonstandard covariates are actually relevant for predicting his on-the-job ability, and all of these nonstandard covariates can be shared during a job interview, then our main results do not extend and we should think of the value of context as being potentially large. On the other hand, if the set of relevant covariates are low-dimensional relative to the total number of covariates, but are still too numerous to be fully revealed, then our main results do extend.

This result suggests that whether human or black box evaluation is more appropriate should be determined in part based on whether the available signal is concentrated in a small number of covariates (favoring the human evaluator) or spread out across a large number of covariates (favoring the black box evaluator). The same application may transition between these regimes over time. For example, in a medical setting where black box diagnosis is highly accurate based on non-interpretable features of an image scan, it may not be possible to communicate sufficient information via any small number of covariates. But if the predictive features of the image are subsequently better understood and defined, then it may be that a small set of (newly defined) features does eventually capture all of the signal content, and can be fully disclosed in a coversation.

5.2 Known effect of certain covariates

Another possibility is that the agent knows how certain nonstandard covariates are correlated with the type.

Example 15. The agent is a patient who resided around Chernobyl at the time of the Chernobyl nuclear disaster of 1986. The agent is being evaluated for potential thyroid conditions, and knows that this particular part of his history increases the probability of a thyroid condition.

Specifically suppose there is a set $K \subseteq \{1, \ldots, n\}$ of covariate indices whose effects are known. The set K includes the standard covariates, but possibly also includes some nonstandard covariates. We weaken Assumption 1 to the following:

Assumption 5. *Fix any realization of the covariates $\mathbf{x}_K = (x_i)_{i \in K}$. Then there is an infinitely exchangeable sequence $(\widetilde{Y}_1, \widetilde{Y}_2, \ldots)$ such that for every $n \in \mathbb{N}$, the sequence*

$$\left(Y_{\mathbf{x}_K, \mathbf{x}_{-K}} : \mathbf{x}_{-K} \in \{0,1\}^{n-|K|}\right)$$

has the same distribution as $(\widetilde{Y}_1, \widetilde{Y}_2, \dots)$.

This assumption imposes exchangeability only over the nonstandard covariates whose effects are not ex-ante known. Clearly if K is a strict superset of $\mathcal{S}$, then the expected value of context need not vanish. A simple example is the following.

Example 16. Suppose there are no standard covariates, and $K = \{1\}$, i.e., the first nonstandard covariate has a known effect, where $f(\mathbf{x}_n) \sim U[-1, 0]$ if $x_1 = 0$ and $f(\mathbf{x}_n) \sim U[0, 1]$ if $x_1 = 1$. Suppose further that the agent's covariate vector satisfies $x_1 = 1$. Then the prior expectation of the agent's type is 0, but revealing $x_1 = 1$ moves the posterior expectation to 1/2. So the expected value of context does not vanish.

But if we modify the definition in (2) to

$$\hat{y}^f_{\mathbf{x}_n}(A) = \mathbb{E}[Y \mid X_i = x_i \quad \forall i \in K \cup A]$$

with K replacing $\mathcal{S}$, and again let $U^f_{\mathbf{x}_n}(A) = u\left(\hat{y}^f_{\mathbf{x}_n}(A), y\right)$, then the modified expected value of context

$$v(f, \mathbf{x}_n) = \max_{\substack{H \subseteq \mathcal{N} \setminus K \\ |H| \leq \alpha_h n}} U^f_{\mathbf{x}_n}(H) - U^f_{\mathbf{x}_n}(\varnothing)$$

evaluates the value of context beyond those covariates with known effects. The same proof shows that this expected value of context vanishes to zero as n grows large. That is, beyond the value of context that is already clear to the agent based on private knowledge about his nonstandard covariates, the agent does not expect substantial additional gain from the remaining covariates.

5.3 Sufficient residual uncertainty

In this final section, 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 $H \subseteq \{s+1, \dots, n\}$ 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 type functions $f : \{0, 1\}^n \to [-\overline{y}, \overline{y}]$, 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

$$Z^n_d = \rho_f(d)$$

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^n_d 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.[28]

Assumption 6 (Unbiased)**.** $\mathbb{E}[Z^n_d] = \mu$ *for every disclosure* d.

Assumption 7 (Fast Concentration)**.** *For any sequence of feasible disclosures* $(d_n)_{n\geq 1}$*,*

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

where $K_n = \sum_{j=0}^{\lfloor \alpha_h n \rfloor} \binom{n-s}{j}$ *is the number of unique sets* $H \subseteq \{s+1, \ldots, n\}$ *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 ρ 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^n_d for any disclosure $d = (H, (x_i)_{i\in H})$ could be represented as a sample average consisting of $2^{n-s-|H|}$ elements. Assumption 6 is clearly satisfied (because the update rule is Bayesian), but one can select a sequence of disclosures (d_n) such that $Var(Z^n_{d_n}) = \frac{1}{2^{n(1-\alpha_h)-s}}$ (see the proof of Theorem 1 for details). Thus the speed of convergence demanded in Assumption 7 is not met when α_h is sufficiently large.

Nevertheless, Assumption 7 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 3. *Suppose Assumptions 6 and 7 hold. Then for every covariate vector* $\mathbf{x} \in \{0,1\}^\infty$*, the expected value of context vanishes to zero as* n *grows large, i.e.,*

$$\lim_{n\to\infty} V(n, \mathbf{x}_n) = 0.$$

This result also clarifies that neither the precise symmetry imposed by Assumption 1, nor the assumption of Bayesian updating in our main model, are crucial for our main result.

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

6 Conclusion

One argument against replacing human experts with algorithmic predictions is that no matter how many covariates are taken as input by the algorithm, the number 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 a black box algorithm.

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 how these nonstandard covariates impact the type, 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 "low-dimensional" set of covariates that predict the type and can be fully disclosed (as in Example 9), or if there is a known structural relationship between covariates and the type (as in Example 11), then the expected value to disclosing additional information may be large. We thus view our results as revealing a link betwen the value of targeting information acquisition (beyond simply conditioning on large quantities of information) and the extent of prior "structural information" about the numerous covariates that can be brought up as explanations.

We conclude with two alternative interpretations of our model and results.

Online versus offline learning. In our model, a key distinction between human and black box evaluation is that the human can adapt which covariates are acquired based on other properties of the agent, while the black box cannot. This is an appropriate comparison of human and black box evaluators as they currently stand: The black box algorithms used to make predictions about humans are usually supervised machine learning algorithms which are pre-trained on a large data set. But new black box algorithms, such as LLMs, 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, how valuable is it to have the evaluator dynamically acquire information given feedback from the agent? Our result suggests that this is not important in expectation. For example, Part (a) of Theorem 2 implies that an agent who cares about accuracy should prefer a supervised machine learning algorithm trained on a large number of covariates over a conversation with ChatGPT that reveals a smaller number of covariates.

Value of human supervision of algorithms. While we have interpreted the s standard covariates as a small set of covariates acquired by the human evaluator, an alternative interpretation is that they are the initial inputs to an algorithm. In this case, the expected value of context quantifies the sensitivity of the algorithm's predictions to the addition of further relevant inputs, e.g., as identified by a human manager. This interpretation is particularly relevant when we consider accuracy as the objective, in which case the value of context tells us how wrong the algorithm is compared to if the algorithm could be retrained on additional relevant inputs. Theorem 1 says that while in certain cases additional inputs would lead to a substantially more accurate prediction, under our symmetry assumption on the agent's prior this will not typically be the case.

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.

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.

Bastani, H., O. Bastani, and W. P. Sinchaisri (2022): "Improving Human Decision-Making with Machine Learning," .

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.

Cripps, M., J. Ely, G. Mailath, and L. Samuelson (2008): "Common Learning," *Econometrica*, 76, 909–933.

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.

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.

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.

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

Jha, S. (2020): "Can you sue an algorithm for malpractice? It depends," .

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.

Kushilevitz, E. and N. Nisan (1996): *Communication Complexity*, Cambridge University Press.

Lai, V., C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan (2023): "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: Association for Computing Machinery, FAccT '23, 1369–1385.

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.

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.

A Proof of Generalization of Theorem 1

In a change of notation relative to the main text, we subsequently use $\mathbf{X}_n$ to denote the agent's covariate vector and Y to denote the agent's type (leaving $\mathbf{x}_n$ and y to denote realizations of these random variables). Moreover, rather than supposing that Y is deterministically related to $\mathbf{X}_n$ via a function f, let $(\mathbf{X}_n, Y) \sim P^n$ where P^n is unknown. We replace Assumption 1 with the following.

Assumption 8. *Fix any realization of the standard covariates* $\mathbf{x}_{\mathcal{S}} \in \{0,1\}^s$. *There is an infinitely exchangeable sequence* $(\widetilde{Y}_1, \widetilde{Y}_2, \dots)$ *such that for every* $n \in \mathbb{N}$, *the sequence*

$$(\mathbb{E}[Y \mid (X_1, \dots, X_n) = (\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}})])_{\mathbf{x}_{-\mathcal{S}} \in \{0,1\}^{n-s}}$$

has the same distribution as $(\widetilde{Y}_1, \dots, \widetilde{Y}_{2^n})$.

That is, permuting the labels and/or values of the nonstandard covariates does not change the joint distribution of the conditional expectations of y. When y is degenerate conditional on $\mathbf{x}_n$, Assumption 8 reduces to our previous Assumption 1. We will prove the following generalization of Theorem 1.

Theorem A.1. *Suppose Assumption 8 holds. Then for every covariate vector* $\mathbf{x} \in \{0,1\}^\infty$, *the expected value of context vanishes to zero as n grows large, i.e.,* $\lim_{n\to\infty} V(n, \mathbf{x}_n) = 0$.

Towards this, we will first prove the conclusion under a strengthening of Assumption 8, where exchangeability is replaced by an assumption that conditional expectations are i.i.d. across the different possible completions of the agent's covariate vector.

Assumption 9. *Fix any realization of the standard covariates* $\mathbf{x}_{\mathcal{S}} \in \{0,1\}^s$. *Then there is a distribution F such that for every* $n \in \mathbb{N}$, *the conditional expectations*

$$\mathbb{E}[Y \mid (X_1, \dots, X_n) = (\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{-\mathcal{S}})] \sim_{iid} F$$

across all vectors $\mathbf{x}_{-\mathcal{S}} \in \{0,1\}^n$.

Theorem A.2. *Suppose Assumption 9 holds. Then for every covariate vector* $\mathbf{x} \in \{0,1\}^\infty$, *the expected value of context vanishes to zero as n grows large, i.e.,* $\lim_{n\to\infty} V(n, \mathbf{x}_n) = 0$.

Sections A.1-A.4 prove Theorem A.2, and Section A.5 shows that Theorem A.2 implies Theorem A.1.

A.1 Outline for Proof of Theorem A.2

Fix any realization $(x_1, \ldots, x_s)$ of the agent's standard covariates. After observing $(x_1, \ldots, x_s)$, the evaluator assigns positive probability to the 2^{n-s} covariate vectors whose first s entries are equal to $(x_1, \ldots, x_s)$. Let these covariate vectors be indexed by $\mathbf{x}^j$ where $j = 1, \ldots, 2^{n-s}$, and define

$$Y_j \equiv \mathbb{E}_{P^n}\left[Y \mid (X_1, \ldots, X_n) = \mathbf{x}^j\right]$$

to be the (random) expected type given covariate vector $\mathbf{x}^j$. By assumption that the marginal distribution over covariate vectors is uniform, the evaluator's posterior expectation of the agent's type after observing the agent's standard covariates is

$$\widehat{Y}(\varnothing, \mathbf{x}_n) = \frac{1}{2^{n-s}} \sum_{i=1}^{2^{n-s}} Y_j \equiv Z^n_{\varnothing}.$$

There are $K_n = \sum_{k=0}^{h_n} \binom{n-s}{k}$ subsets of $\{s+1, \ldots, n\}$ that contain h_n or fewer elements. Enumerate these sets as $H_1, \ldots, H_{K_n}$. For each H_k, let

$$S_k = \left\{j \,:\, \mathbf{x}^j \in C_{H_k}(\mathbf{x}_n)\right\}$$

be the set of indices for those covariate vectors $\mathbf{x}^j$ that agree with the agent's covariate vector $\mathbf{x}_n$ in entries $(1, \ldots, s) \cup H_k$ (where $C_{H_k}(\mathbf{x}_n)$ is as defined in (1)). After observing the agent's nonstandard covariates in the set H_k, the evaluator's posterior expectation about the agent's type is

$$\widehat{Y}(H_k, \mathbf{x}_n) = \frac{\sum_{j \in S_k} Y_j}{|S_k|} \equiv Z_k.$$

Although the distributions of the random variables Z_k vary across n, we suppress this dependence in what follows to save on notation. The remainder of the proof proceeds by first showing that in expectation the possible increase in the evaluator's posterior expectation over the prior expectation $\mu \equiv \mathbb{E}[Y]$ is vanishing.

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

This is subsequently strengthened to the statement that the expected maximum absolute difference between Z_k and μ converges to zero.

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

And finally we apply the above proposition to demonstrate the conclusion of the theorem, i.e., that

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

Thus in expectation the possible increase in the agent's payoff also vanishes. We suppress dependence of V on the covariate vector $\mathbf{x}_n$ in what follows, writing simply $V(n)$.

A.2 Proof of Proposition A.1

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

The quantity $\mathbb{E}[\max_{1\le k\le K_n} Z_k]$ 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.2.1 and A.2.2, we define i.i.d. variables Z_k^{iid} with the property that

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

In Sections A.2.3 and A.2.4, we show that the RHS of the above display converges to μ as n grows large.

A.2.1 Replacing Z_k's with independent variables Z_k^{ind}

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 $1 \le k \le K_n$ define*

$$Z_k^{ind} = \frac{\sum_{j=1}^{|S_k|} Y_j^k}{|S_k|} \tag{A.2}$$

where $Y_j^k \sim_{iid} F$, so that each Z_k^{ind} has the same distribution as Z_k, but the vector $(Z_1^{ind}, \ldots, Z_K^{ind})$ 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 $V_n \le V_n^{ind}$ 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*

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

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

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

$$\mathbb{P}(M_C^q \le t) - \mathbb{P}(\widetilde{M}_C^q \le t) \ge 0$$

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

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

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

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

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 $X_1 + Y_q \le \max\{X_i + a_i Y_1\}_{i=1}^{q-1}$, 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. $\square$

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 $q \in \{1, \ldots, Q-1\}$,

$$\begin{aligned} \widehat{M}_C^q &= \max\{M_C^q, W^q\} \\ \widehat{M}_C^{q-1} &= \max\{\widetilde{M}_C^q, W^q\} \end{aligned}$$

where $W^q = \max\left\{\max\{X_i + a_i Y_i\}_{i=q}^Q, W\right\}$ 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. $\square$

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

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

or

$$\max\left\{\max_{k\in I}\left\{X_k + \frac{1}{|S_k|}Y_i\right\}, W\right\} \tag{A.3}$$

where $X_k \equiv \frac{1}{|S_k|}\sum_{j\in S_k, j\neq i} Y_j$ for each $k \in I$, and $W \equiv \max_{k\notin I} \frac{\sum_{j\in S_k} Y_k}{|S_k|}$. 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.3) 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.2.2 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-s-h_n} and 2^{n-s} elements. We next show that replacing each of these variables with a sample average of 2^{n-s-h_n} elements (the smallest size) weakly increases the value of context.

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

$$Z_k^{iid} = \frac{\sum_{j=1}^{2^{n-s-h_n}} Y_j^k}{2^{n-s-h_n}} \tag{A.4}$$

to be the analogue of Z_k^{ind} with 2^{n-s-h_n} elements instead of $|S_k| \geq 2^{n-s-h_n}$, 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 $|S_k| \geq 2^{n-s-h_n}$ for all k). The desired result then follows by Jensen's inequality, since the entries of $(Z_1^{ind}, \ldots, Z_K^{ind})$ are (by construction) independent and the maximum is a convex function. □

A.2.3 Asymptotic Normality

Lemma A.3. *Let*

$$V_n^N \equiv \mathbb{E}\left[\max\{Z_1^N, \ldots, Z_{K_n}^N\}\right]$$

where $Z_k^N \sim \mathcal{N}\left(\mu, \frac{1}{2^{n-s-h_n}}\right)$. *Then* $\lim_{n\to\infty} |V_n^{iid} - V_n^N| = 0$.

Proof. Without loss of generality, let $Var(Y_j^k) = 1$.[29] First observe that

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

where each

$$\widetilde{Z}_k^{iid} = \frac{1}{\sqrt{2^{n-s-h_n}}} \sum_{i=1}^{2^{n-s-h_n}} Y_i^k.$$

Similarly we can write

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

where each

$$\widetilde{Z}^N \sim_{iid} \mathcal{N}(\mu, 1).$$

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

When the assumptions for Corollary 2.1 from Chernozhukov et al. (2013) are met (to be verified momentarily), we can conclude that

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

where ρ denotes Kolmogorov distance. Thus also

$$\rho(M_n^{iid}, M_n^N) \to 0 \tag{A.5}$$

where

$$\begin{aligned} M_n^{iid} &= \frac{1}{\sqrt{2^{n-s-h_n}}} \max\{\widetilde{Z}_1^{iid}, \dots, \widetilde{Z}_{K_n}^{iid}\} \\ M_n^{N} &= \frac{1}{\sqrt{2^{n-s-h_n}}} \max\{\widetilde{Z}_1^{N}, \dots, \widetilde{Z}_{K_n}^{N}\} \end{aligned}$$

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.5) 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

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

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

A.2.4 Upper Bound for Expected Maximum of Gaussians

Finally by Berman (1964), which provides an upper bound for the expected maximum of independent Gaussian random variables, there exists a positive constant C such that

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

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

A.3 Proof of Proposition A.2

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

In an abuse of notation, let $Z_k \equiv Z_k - \mu$ denote de-meaned sample average. By rewriting the max within the expectation we obtain

$$\mathbb{E}\left[\max_{1\le k\le K_n} |Z_k|\right] = \mathbb{E}\left[\max\left\{\max_{1\le k\le K_n} Z_k, -\min_{1\le k\le K_n} Z_k\right\}\right]$$
$$\le \mathbb{E}\left[\max\left\{\max_{1\le k\le K_n}\{Z_k\}, 0\right\}\right] + \mathbb{E}\left[\max\left\{-\min_{1\le k\le K_n}\{Z_k\}, 0\right\}\right]$$

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

$$\mathbb{E}\left[\max\left\{\max_{1\le k\le K_n}\{Z_k\}, 0\right\}\right] = \mathbb{P}\left(\max_{1\le k\le K_n} Z_k \ge 0\right)\cdot \mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k \ge 0\right] \tag{A.6}$$

Moreover,

$$\mathbb{E}\left[\max_{1\le k\le K_n} Z_k\right] = \mathbb{P}\left(\max_{1\le k\le K_n} Z_k \ge 0\right)\cdot \mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k \ge 0\right]$$
$$+ \mathbb{P}\left(\max_{1\le k\le K_n} Z_k < 0\right)\cdot \mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k < 0\right]$$

so

$$\mathbb{P}\left(\max_{1\le k\le K_n} Z_k \ge 0\right)\cdot \mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k \ge 0\right] =$$
$$= \mathbb{E}\left[\max_{1\le k\le K_n} Z_k\right] - \mathbb{P}\left(\max_{1\le k\le K_n} Z_k < 0\right)\cdot \mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k < 0\right] \tag{A.7}$$

From Lemma A.1,

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\le k\le K_n} Z_k\right] = 0. \tag{A.8}$$

Moreover, we showed in Section A.2.1 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

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

which converges to zero as n grows large since each $\mathbb{P}(Z_k^{ind} < 0) < 1$. Finally,

$$\mathbb{E}\left[\max_{1\le k\le K_n} Z_k \mid \max_{1\le k\le K_n} Z_k < 0\right] \in \left[-\overline{Y}, \overline{Y}\right] \tag{A.9}$$

uniformly across n. Putting together (A.6) - (A.9) we have that

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

as desired. The argument that

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

follows identically, observing that Proposition A.1 is satisfied for $\widetilde{Y} \equiv -Y$, and that

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

A.4 Concluding the proof of Theorem A.2

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

$$\begin{aligned} V(n) &= \mathbb{E}\left[\max_{1\le k\le K_n} u(Z_k, Y)\right] - \mathbb{E}\left[u(Z_\varnothing^n, Y)\right] \\ &\le \mathbb{E}\left[\max_{1\le k\le K_n} |u(Z_k, Y) - u(Z_\varnothing^n, Y)|\right] \end{aligned} \tag{A.10}$$

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

$$|u(Z_k, Y) - u(Z_\varnothing^n, Y)| \le |u(Z_k, Y) - u(\mu, Y)| + |u(\mu, Y) - u(Z_\varnothing^n, Y)| \tag{A.11}$$

Since u is by assumption Lipschitz continuous in the first argument, there is a constant B such that

$$|u(z_k, y) - u(\mu, y)| \le B|z_k - \mu| \tag{A.12}$$

and

$$|u(\mu, y) - u(z_\varnothing, y)| \le B|z_\varnothing - \mu| \tag{A.13}$$

for any realizations z_k and $z_\varnothing$ of Z_k and $Z_\varnothing^n$. Combining equations A.10-A.13 we get

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

Clearly $\mathbb{E}[Z_\varnothing^n] = \mu$. Moreover, by assumption that each Y is uniformly bounded above and below, the sequence $(Z_\varnothing^n)$ is uniformly integrable. It follows from the Law of Large Numbers that

$$\lim_{n\to\infty} \mathbb{E}[|Z_\varnothing^n - \mu|] = 0$$

Finally, $\lim_{n\to\infty} \mathbb{E}\left[\max_{1\le k\le K_n} |Z_k - \mu|\right] = 0$ follows directly from Lemma A.2. So the RHS of A.11 converges to zero, implying $V(n) \to 0$ as desired.

A.5 Theorem A.2 implies Theorem A.1

In an abuse of notation, let $P^n \sim F$ mean that $Y_{\mathbf{x}_n} \sim_{iid} F$ across all covariate vectors $\mathbf{x}_n$. We have already shown in Theorem 1 that $\lim_{n\to\infty} \mathbb{E}_{P^n\sim F}(v_n(P)) = 0$ for any distribution F. Now suppose instead that Assumption 1 is satisfied. By de Finetti's theorem, there exists a set Θ, family of conditional measures $(\pi_\theta)_{\theta\in\Theta}$, and measure $\nu \in \Delta(\Theta)$ such that

$$V(n, \mathbf{x}) = \int_\Theta \mathbb{E}_{P^n\sim F_\theta}(v_n(P, \mathbf{x}_n))d\nu(\theta)$$

where the inner expectation converges to zero for every θ by Theorem A.2. By assumption that u is Lipschitz continuous on a compact domain, there exist $\underline{u}$ and $\overline{u}$ such that $u(\hat{y}, y) \in [\underline{u}, \overline{u}]$ for all $(\hat{y}, y)$. So $\mathbb{E}_{P^n\sim F_\theta}(v_n(P, \mathbf{x}_n))$ is pointwise bounded above by $\overline{u} - \underline{u}$, and we can apply the Dominated Convergence Theorem to conclude that $\lim_{n\to\infty} V(n, \mathbf{x}) = 0$ as desired.

A.6 Proof of Theorem 2

Throughout the proof we set $s = 0$, $\mu = 0$ and $\sigma^2 = \mathbb{E}(Y_i^2) = 1$ without loss of generality. We'll start by demonstrating Part (a). As before let $B_n \subseteq \{1, \ldots, 2^n\}$ index those 2^{n-b_n} covariate vectors that agree with the agent's covariate vector for all covariates in B. Then the black box evaluator's posterior expectation is the sample average

$$Z_B^n = \frac{1}{2^{n-b_n}} \sum_{j\in B_n} Y_j.$$

We will show that

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

for large enough n.

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

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

for some $\tilde{Z} \in [0, Z_B^n]$. Note that $\mathbb{E}[Z_B^n] = \mathbb{E}[Y] = 0$. Moreover, $\phi''(\tilde{Z}) \ge c_1 > 0$ for some c_1, since ϕ is strictly convex. Thus

$$\mathbb{E}[\phi(Z_B^n) - \phi(0)] \geq c_1 \mathbb{E}[(Z_B^n)^2] = \frac{c_1}{2^{(1-\alpha_b)n}}$$

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

$$\phi(Z_k) - \phi(0) \leq c_2 |Z_k|$$

where the latter inequality follows from the fact that ϕ' is continuous on a compact set, and hence bounded by some $c_2 \geq 0$. Thus

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

As a corollary of Lemma A.2 it follows that there exist $c_3, c_4 > 0$ such that

$$\mathbb{E}[\max\{|Z_k|\}] \leq c_3 \frac{1}{2^{(1-\alpha_h)n}} \sqrt{\log(c_4 K_n)}$$

And, thus

$$\mathbb{E}\left[\max_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] \leq c_2 c_3 \frac{1}{2^{(1-\alpha_h)n}} \sqrt{\log(c_4 K_n)}$$

Combining the bounds from steps 1 and 2 we get

$$\Delta(n) \geq c_1 \frac{1}{2^{(1-\alpha_b)n}} - c_2 c_3 \frac{1}{2^{(1-\alpha_h)n}} \sqrt{\log(c_4 K_n)}$$

The RHS is asymptotically positive if and only if

$$\frac{2^{(1-\alpha_h)n}}{2^{(1-\alpha_b)n}} \xrightarrow{n\to\infty} \infty$$

since $\sqrt{\log(c_4 K_n)}$ has sub-exponential but non-constant asymptotics. This condition is satisfied if and only if $\alpha_b > \alpha_h$.

Part (b) follows by identical arguments: Since $-\phi$ is convex, the above arguments apply to show that

$$\mathbb{E}[\phi(Z_B^n) - \phi(0)] \leq -\frac{c_1}{2^{(1-\alpha_b)n}}$$

for some $c_1 > 0$, while

$$\begin{aligned}\mathbb{E}\left[\min_{1\leq k\leq K_n} \phi(Z_k) - \phi(0)\right] &= -\mathbb{E}\left[\max_{1\leq k\leq K_n} -\phi(Z_k) - (-\phi(0))\right] \\ &\geq -c_2 c_3 \frac{1}{2^{(1-\alpha_h)n}} \sqrt{\log(c_4 K_n)}\end{aligned}$$

for some $c_2, c_3, c_4 > 0$. The desired conclusion follows.

A.7 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 A.3. *There exists an N sufficiently large such that the agent prefers the black box evaluator for all $n \geq N$.*

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

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

for large enough n.

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

$$\mathbb{E}[u(Z_B^n, y) - u(0,y)] = \mathbb{E}\left[(Z_B^n)^2\right] = \frac{1}{2^{(1-\alpha_B)n}}.$$

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 c_2 such that $u(z_k, y) - u(0,y) \leq c_2|z_k|$ holds pointwise for each realization 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. □

B Proofs for Results in Sections 4 and 5

B.1 Proof of Corollary 1

We continue in the general setting outlined in the proof of Theorem A.1. Fix any realization $\mathbf{x}_{\mathcal{S}} = (x_1, \ldots, x_s)$ of the standard covariates. As in the proof of Theorem 1, there are 2^{n-s}

covariate vectors $\mathbf{x}_n \in \{0,1\}^n$ with positive probability conditional on $\mathbf{x}_{\mathcal{S}}$. Index these by $j = 1, \ldots, 2^{n-s}$, and define

$$Y_j^{\mathbf{x}_{\mathcal{S}}} \equiv \mathbb{E}_{P^n}\left[Y \mid (X_1, \ldots, X_n) = \mathbf{x}_n^j\right]$$

to be the expected type given covariate vector $\mathbf{x}_n^j$. For each covariate vector $\mathbf{x}_n$ and each disclosure set $D_k \subseteq \{s+1, \ldots, n\}$, there is a corresponding set of covariate vectors S_k such that the evaluator's posterior expectation after the agent discloses his covariates in set D_k is

$$Z_k^{\mathbf{x}_{\mathcal{S}}} = \frac{\sum_{j \in S_k} Y_j^{\mathbf{x}_{\mathcal{S}}}}{|S_k|}.$$

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

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

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

$$\mathbb{E}\left(\max_{1 \leq k \leq \overline{K}_n} |Z_k^{\mathbf{x}_{\mathcal{S}}} - \mu|\right) \leq \frac{1}{2^{n-s-h_n}} C \sqrt{\log(\overline{K}_n)} \leq \frac{1}{2^{n(1-\alpha_h)-s}} C \sqrt{\log(3^{n-s})}$$

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

$$\begin{aligned}\mathbb{E}\left[\max_{\mathbf{x}_{\mathcal{S}} \in \{0,1\}^s} \left(\max_{1 \leq k \leq K_n} |Z_k^{\mathbf{x}_{\mathcal{S}}} - \mu|\right)\right] &\leq \mathbb{E}\left[\sum_{\mathbf{x}_{\mathcal{S}} \in \{0,1\}^s} \max_{1 \leq k \leq K_n} |Z_k^{\mathbf{x}_{\mathcal{S}}} - \mu|\right] \\ &= \sum_{\mathbf{x}_{\mathcal{S}} \in \{0,1\}^s} \mathbb{E}\left[\max_{1 \leq k \leq K_n} |Z_k^{\mathbf{x}_{\mathcal{S}}} - \mu|\right].\end{aligned}$$

Since each $\mathbb{E}\left[\max_{1 \leq k \leq K_n} |Z_k^{\mathbf{x}_{\mathcal{S}}}|\right] \to 0$ 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.

B.2 Proof of Proposition 1

Throughout this proof, we set $s = 0$ for simplicity of notation.

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

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

Proof. Consider a candidate equilibrium (σ^*, μ^*), where $\sigma^*(\mathbf{x}_n) = \varnothing$ for all $\mathbf{x}_n \in \{0,1\}^n$ (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 out-of-equilibrium beliefs such that $u(Z_{\mu*}(\varnothing)) \geq u(Z_{\mu^*}(d))$. This is always possible, for example by setting $Z_{\mu*}(\varnothing) = Z_{\mu^*}(d)$ 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^*, μ^*) of the f-context disclosure game. Let $d_1, \ldots, d_N$ index the disclosures that have positive probability under σ^* (i.e., all $d \in \mathcal{D}$ such that $\sigma^*(\mathbf{x}_n) = d$ for some $\mathbf{x}_n$). For each such disclosure d_i,

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

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 $Z_{\mu^*}(d_i)$. Let

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

be the disclosure that yields the highest payoff to the Sender. Then it must be that $\sigma^*(\mathbf{x}_n) = d^*$ for every covariate vector $\mathbf{x}_n$ 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

$$u(Z_{\mu^*}(d^*)) \leq \max_{\mathbf{x}_n \in \{0,1\}^n} v(f, \mathbf{x}_n).$$

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

B.3 Result for Mixed Strategy Equilibria

In this part we restrict to equilibria (σ^*, μ^*) with the property that $\arg\max_{\hat{y} \in A_{(\sigma^*,\mu^*)}} u(\hat{y})$ is unique on the set $A_{(\sigma^*,\mu^*)}$ 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 $v^D(f, \mathbf{x}_n)$ denote the highest payoff that an agent with covariate vector $\mathbf{x}_n$ receives in any generic equilibrium (potentially mixed) of the f-context disclosure game. Further define

$$v_f^D(n) = \max_{\mathbf{x}_n} v^D(f, \mathbf{x}_n)$$

and

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

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

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

Proof. Fix n, f, and a context equilibrium (σ^*, μ^*) 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

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

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 $\mathcal{X}^* \subseteq \{0,1\}^n$ 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

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

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 $\mathcal{X}^*$.[30]

Now, consider the following alternative "strategy" σ_0, which relaxes the feasibility constraint: For any $\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}^*$ 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 $\mathbf{x} \in \mathcal{X}^*$. 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 (B.2), we can write

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

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

Combining equations (B.2) and (B.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 $C_d \subseteq \{0,1\}^n$ 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.

□

B.4 Proof of Proposition 2

We again continue in the general setting outlined in the proof of Theorem A.1, and adopt the conventions that $\mathbb{E}(Y) = \mu$ while $\text{Var}(Y) = 1$. Recalling that r_n is the number of relevant covariates, there are 2^{r_n} distinct expected conditional types, which we can enumerate as $Y_1, \ldots, Y_{2^{r_n}}$. If disclosure k involves disclosing k_r relevant covariates, then there is a set S_k of size $2^{r_n - k_r}$ such that the evaluator's posterior expectation can be written

$$Z_k = \frac{1}{2^{n-h_n}} \sum_{j\in S_k} 2^{n-r_n-(h_n-k_r)} Y_j = \frac{1}{2^{r_n-k_r}} \sum_{j\in S_k} Y_j.$$

As in Step 1 of the proof of Theorem 1 (Section A.2.1), replace each Y_j with a variable $Y_j^k \overset{d}{=} Y_j$ which is independent across disclosure sets. This yields the random variables

$$Z_k^{ind} = \frac{1}{2^{r_n-k_r}} \sum_{j\in S_k} Y_j^k.$$

As in the proof of Proposition A.1, it follows from Lemma 1 that

$$\mathbb{E}[\max\{Z_1, \ldots, Z_{K_n}\}] \leq \mathbb{E}[\max\{Z_1^{ind}, \ldots, Z_{K_n}^{ind}\}].$$

Next define

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

and note that these are identically and independently distributed with shared variance

$$Var(Z_k^{iid}) = \frac{1}{2^{r_n - h_n}}.$$

Following the arguments in Step 2 of the proof of Theorem 1 (Section A.2.2), we get

$$\mathbb{E}[\max\{Z_1^{ind}, \ldots, Z_{K_n}^{ind}\}] \leq \mathbb{E}[\max\{Z_1^{iid}, \ldots, Z_{K_n}^{iid}\}].$$

where as before $K_n = \sum_{j=0}^{h_n} \binom{n}{j}$. Further, by the argument given in Step 3 of the proof of Theorem 1 (Section A.2.3),

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

where

$$V_n^N \equiv \mathbb{E}\left[\max\{Z_1^N, \ldots, Z_{K_n}^N\}\right]$$

and $Z_k \sim \mathcal{N}\left(\mu, \frac{1}{2^{r_n - h_n}}\right)$. Again applying the bound from Berman (1964), we have

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

By assumption that $\alpha_r > \alpha_h$, the right-hand expression converges to zero as n grows large, concluding the proof.

B.5 Proof of Proposition 3

Throughout the proof we assume $u(x) \equiv x$ and $s = 0$. In addition, for simplicity of notation, we enumerate feasible disclosures by k and denote the corresponding posteriors (as random variables) as $Z_k^n := \rho_f(d_k)$. To upper bound the value of context, we apply a result from Arnold and Groeneveld (1979):

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

By Assumption 6, inequality B.4 simplifies to

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

Finally, Assumption 7 implies that $Var(Z_k^n) = o(\frac{1}{K_n})$ for every disclosure k. Hence

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

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.