DATA-DRIVEN MECHANISM DESIGN: JOINTLY ELICITING PREFERENCES AND INFORMATION

By

Dirk Bergemann, Marek Bojko, Paul Dütting, Renato Paes

Leme, Haifeng Xu and Song Zuo

March 2025

COWLES FOUNDATION DISCUSSION PAPER NO. 2418R1

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Data-Driven Mechanism Design: Jointly Eliciting Preferences and Information

Dirk Bergemann* Marek Bojko† Paul Dütting‡

Renato Paes Leme§ Haifeng Xu¶ Song Zuo‖

March 17, 2025

Abstract

We study mechanism design when agents have private preferences and private information about a common payoff-relevant state. We show that standard message-driven mechanisms cannot implement socially efficient allocations when agents have multidimensional types, even under favorable conditions.

To overcome this limitation, we propose data-driven mechanisms that leverage additional post-allocation information, modeled as an estimator of the payoff-relevant state. Our data-driven mechanisms extend the classic Vickrey-Clarke-Groves class. We show that they achieve exact implementation in posterior equilibrium when the state is either fully revealed or the utility is affine in an unbiased estimator. We also show that they achieve approximate implementation with a consistent estimator, converging to exact implementation as the estimator converges, and present bounds on the convergence rate.

We demonstrate applications to digital advertising auctions and large language model (LLM)-based mechanisms, where user engagement naturally reveals relevant information.

Keywords: Mechanism Design, Data-Driven Mechanism Design, Large Language Models, Click-Through Rate, Posterior Equilibrium

JEL Codes: D47, D82, D83

*Department of Economics, Yale University, dirk.bergemann@yale.edu

†Department of Economics, Yale University, marek.bojko@yale.edu

‡Google Research, duetting@google.com

§Google Research, renatoppl@google.com

¶Department of Computer Science, University of Chicago and Google Research, haifengxu@uchicago.edu

‖Google Research, szuo@google.com

*Dirk Bergemann gratefully acknowledges financial support from NSF SES 1948336 and SES 2049754 and ONR MURI. Haifeng Xu acknowledges financial support from NSF Award CCF-2303372 and AI2050 award G-24-66104 of Schmidt Sciences. We thank Luis Hoderlein, Elchanan Mossel, Andrew Postlewaite and participants of the Yale Micro-Theory Breakfast seminar for helpful comments. All errors are our own.

1 Introduction

1.1 Motivation

The rise of data-rich digital environments has created new opportunities and challenges for mechanism design. In settings ranging from online advertising to large language models (LLMs), participants often possess private information not just about their preferences but also about underlying states that affect all agents' payoffs. For instance, in sponsored search auctions, advertisers have private information about both their value per click and user behavior patterns. With the emerging LLM technology, content providers increasingly use LLMs to generate content; they have private knowledge of both their desired output and the preferences of downstream consumers.[1] Efficiently aggregating and incentivizing the revelation of this multi-dimensional private information is crucial for the optimal allocation. While the tension between disentangling preferences from information in mechanism design and related environments is well-documented,[2] these "modern" applications highlight the timeliness and relevance of revisiting these questions. At the same time, these applications provide an abundance of additional data revealed through user interactions, platform feedback, or other observable outcomes.[3] As we will show, this additional data can be used to reconcile these two forces.

In a static setting with interdependent values and quasi-linear preferences, we develop a framework for mechanism design that moves beyond traditional message-driven approaches by incorporating such naturally available data. Rather than relying solely on agents' reported messages, our mechanisms condition transfers on additional information about the state. This data-driven approach enables the implementation of efficient allocations, maximizing the sum of agents' expected payoffs based on their combined information, even when agents have multi-dimensional private types—a setting where standard mechanisms provably fail. We specifically focus on implementation in posterior equilibrium, ensuring that no agent is incentivized to deviate from truthful reporting, provided that all other agents report truthfully, even after the uncertainty about other agents' types has resolved. Implementation in posterior equilibrium thus ensures robustness to belief misspecification and eliminates incentives for agents to inefficiently acquire information about other agents' types.

1.2 Framework and Results

In our model, agents hold private preferences over allocations and are endowed with private information about a common payoff-relevant state. We model each agent's information about

[1] A prominent example is the novel format of digital advertising auctions, where advertisers bid to feature in "sponsored" content generated by an LLM. At the time of this writing, Perplexity AI, a widely used LLM-based chatbot with an integrated search engine, began incorporating such digital advertising auctions on its platform; see this Financial Times article: https://on.ft.com/47BWbIX.

[2] See, for example, Lu (2019) and references therein for a related issue of identification of beliefs and state-dependent utilities.

[3] For instance, in sponsored search auctions, this additional information includes click-through rates on ads. In settings involving LLMs, it encompasses metrics such as clicks on suggested links, follow-up queries, and engagement duration.

the state as a signal and its realization. We extend the state space to the product probability space of the payoff-relevant state with a common prior and the unit interval with the Lebesgue measure representing residual randomness.[4] A signal is a random variable on the extended state space—a measurable function mapping the extended state to a signal realization from a fixed space of feasible realizations. The formulation allows for arbitrary correlation of agents' signals, an important feature of our motivating examples, as agents might obtain data from similar sources with overlapping datasets. Unlike the literature described in detail below, we do not assume signals are common knowledge. Instead, signals are private and form a component of agents' types. A mechanism elicits signals together with preferences and signal realizations. The formulation reflects practical considerations regarding data sharing. Proper interpretation of the data requires not only access to the dataset but also documentation of the data-generating process—both of which constitute proprietary information.

We begin our analysis with the standard message-driven mechanism design framework, where allocations and transfers depend solely on the messages sent by agents to the mechanism. In interdependent value settings with commonly known signals, Maskin (1992), Dasgupta and Maskin (2000), Jehiel and Moldovanu (2001), and Bergemann and Välimäki (2002) established an implementation impossibility when agents have multi-dimensional types or, in the case of single-dimensional types, when an appropriate single-crossing condition is not met. Given that agents in our model possess private information regarding both the state and their preferences, private types are inherently multi-dimensional. In Theorem 1, we show that implementation in our setting remains impossible even under conditions where previous possibility results emerged: there is an instance with commonly known signals, single-dimensional signal realizations and supermodular expected payoffs for which no message-driven mechanism implements the efficient allocation in posterior equilibrium.

To address the impossibility result, we adopt an approach similar to Mezzetti (2004) by relaxing the restriction that transfers rely solely on the agents' messages. Instead, we allow transfers to be conditioned on additional information about the state, while maintaining that allocations are determined only by messages. In a setting similar to ours, Mezzetti (2004) proposed a two-stage mechanism: in the first stage, the mechanism elicits agents' types to determine the allocation, and in the second stage, it elicits realized payoffs to compute transfers. However, this approach assumes that agents can fully observe and report their final payoffs, and that the second stage is feasible—conditions that may be restrictive or unworkable due to institutional constraints.[5] We show implementation in posterior equilibrium can be obtained under strictly weaker requirements.

[4] We extend the main model to heterogeneous priors over the common payoff-relevant state in Section 4.2.

[5] For instance, Jehiel and Moldovanu (2005) emphasize the fragility of the mechanisms proposed by Mezzetti (2004). This fragility arises because agents may have noisy and subjective perceptions of the state, which influence their payoffs and undermine the mechanism's reliability. Additional challenges, such as moral hazard, verifiability, and other practical concerns, further exacerbate this issue. Notably, in the second stage of the two-stage generalized VCG mechanism proposed by Mezzetti (2004), agents are indifferent to their utility reports, in contrast to the typical strict incentives for truthful reporting seen in standard VCG mechanisms, particularly in private value settings.

We formalize the additional information as an estimator of the payoff-relevant state available to the designer after allocation but before finalizing transfers, with the estimator's data-generating process being common knowledge. The estimator is assumed independent of agents' signals conditional on the payoff-relevant state. In our motivating examples, this approach captures scenarios where the designer gathers data on user interactions both within and beyond the auction environment, thereby gaining insights into user preferences and demand (i.e., the state). As highlighted, such data is naturally available and already collected by platforms, providing a straightforward interpretation for our framework. Notably, unlike Mezzetti (2004), our method eliminates the need for a second reporting stage or access to agents' true preferences, thereby reducing the amount of information necessary for implementation.

We introduce a modified version of Vickrey–Clarke–Groves (VCG) mechanisms (Vickrey, 1961; Clarke, 1971; Groves, 1973), which we call *data-driven VCG mechanisms* (Definition 6). In data-driven VCG transfers, the designer inserts the obtained estimate of the state into agents' payoffs. We then examine how various properties of the estimator affect the feasibility of implementation. We first show that with full resolution of uncertainty, data-driven VCG mechanisms achieve implementation in posterior equilibrium (Theorem 2). Specifically, if the designer ultimately observes the true state, she can utilize ex-post utilities based on the reported types to calculate VCG payments, aligning agents' incentives. This alignment occurs because agents evaluate these transfers according to their true posterior beliefs, making truthful reporting optimal under efficient allocation.

Next, we assume the designer only has access to a noisy estimate of the state and assess the implications of two key properties of estimators common in statistics and econometrics. First, when using an unbiased estimator and assuming agents' utility functions are affine in the state, data-driven VCG mechanisms yield implementation in posterior equilibrium (Proposition 1). However, as anticipated by Jensen's inequality, this result does not hold universally beyond affine utilities. Nevertheless, our application to click-through auctions demonstrates that an unbiased estimator can be beneficial in settings with risk-neutral bidders and appropriately defined states.

Second, we consider the property of consistency, where a sequence of estimators, indexed by the dataset size, converges in probability to the true state as the dataset size grows large. Subject to regularity conditions, Theorem 3 establishes that any corresponding sequence of data-driven VCG mechanisms achieves implementation in ϵ-posterior equilibrium, where no agent has more than an ϵ additive utility loss of having reported truthfully after the resolution of uncertainty about others' types when other agents also report truthfully, with ϵ approaching zero as the sample size increases. That is, while reporting truthfully is an ϵ-posterior equilibrium in any finite sample for large enough ϵ for any estimator, consistent estimators ensure the ϵ can be made arbitrarily small in the limit. Proposition 2 further extends this result by linking the rate of convergence of the estimator to the rate at which ϵ approaches zero. Assuming uniform integrability of the sequence of relevant random variables, we demonstrate that the sequence of ϵ's can be reduced at essentially the same rate as the estimator converges in probability to the true state.

1.3 Applications

We first apply our framework to click-through auctions, focusing on a position auction for a single advertising slot for simplicity. Following the standard specification of agents' payoffs (Edelman et al., 2007; Varian, 2007), we treat agents' values per click as their preference types and the click-through rates as the state. Within this framework, mechanisms based solely on per-impression payments conditioned only on agents' actions are subject to the same impossibility result identified earlier. In contrast, the commonly used per-click payments fall within the class of our data-driven mechanisms. However, their effectiveness in eliciting truthful reporting of both agents' values and signals hinges on whether the click-through rates are common or agent-specific.

When click-through rates are common, per-click pivot payments achieve implementation in posterior equilibrium. These payments align with the data-driven VCG mechanisms class, as the observed frequency of clicks serves as an unbiased (and consistent) estimator of the true click-through rate if clicks are sampled independently and identically (i.i.d.). However, with agent-specific click-through rates, data-driven VCG transfers cannot be feasibly structured as per-click payments, and feasible per-click payment rules become vulnerable to manipulation.

Our second application focuses on mechanism design for content generated by LLMs. Each input prompt defines a distinct mechanism design environment. We model an LLM as a conditional prediction system that selects a generation distribution over feasible output text units to maximize the expected value of a provided reward function, leveraging its training data, knowledge bases, and model architecture—represented as signals and preference types within our framework. We introduce a reference LLM that represents the platform's organic generation distribution, aiming to align responses with user preferences. The designer's objective is to determine a central LLM generation distribution that maximizes the total reward of agents while accounting for a regularization term that measures deviations from the reference distribution. Incorporating this regularization term into each agent's transfer within data-driven VCG mechanisms leads to the definition of regularized data-driven VCG mechanisms (Definition 10). These mechanisms retain the key properties established in our main analysis (Corollary 3). To illustrate, we provide an example where the regularization term is specified as the Kullback-Leibler divergence between the selected distribution and the reference distribution (Example 4).

1.4 Related Literature

This paper contributes to multiple strands of the mechanism design literature, including: (i) efficient mechanism design in settings with interdependent values, (ii) sponsored search auctions, and (iii) mechanism design for LLMs.

The literature on mechanism design with interdependent values, initiated by Milgrom and Weber (1982), is extensive. More recent research (Maskin, 1992; Jehiel and Moldovanu, 2001; Bergemann and Välimäki, 2002) has shown that incentive compatibility and efficiency cannot be simultaneously achieved when the random variables defining signals are commonly known, signals are independent and multidimensional, or single-dimensional if the single-crossing condi-

tion is violated. These impossibility results extend to implementation in posterior equilibrium, irrespective of the stochastic structure of the signals.

Our impossibility result builds on this literature, relying on the assumption that agents have a strictly positive informational size, as defined by McLean and Postlewaite (2002). In contrast, McLean and Postlewaite (2015) demonstrate that when agents have zero informational size, implementation in posterior equilibrium is possible. Furthermore, they establish a continuity result: as agents' informational size approaches zero, approximate implementation becomes feasible, converging to exact implementation in the limit. We also provide a continuity result, albeit through a different channel, while allowing for arbitrary informational sizes of the agents.

Several papers propose two-stage mechanisms. Assuming a common prior on signal realizations and states, McLean and Postlewaite (2017) let agents first report their signal realizations and the designer announce the implied posterior on the state, reducing the problem to a private value setting. In the second stage, agents report their expected payoff functions. Allocations and transfers are determined using a VCG mechanism based on these reports. Mezzetti (2004) considers a framework where the allocation is fixed upfront, but transfers depend on agents' subsequent reports of their final payoffs. The central result establishes that efficient allocation can always be implemented in perfect Bayesian equilibrium via a two-stage VCG mechanism. Instead, we show how additional information often readily available to the designer can be utilized to align incentives without requiring further communication. Our data-driven transfer scheme also addresses practical challenges, such as when agents cannot observe final payoffs or a two-stage mechanism is infeasible. Furthermore, compared to Mezzetti (2004), our results imply that implementation is achievable with strictly less information transmitted to the mechanism: agents' true ex-post payoffs need not be reported for implementation in posterior equilibrium.

Wu et al. (2024) study an auction design setting where bidders also have both private preferences and information. However, for the information component, they only allow bidders to reveal partial information but forbid misreports, whereas our model allows information misreporting.

The work cited above assumes a commonly known prior over the state and agents' signal realizations. Our model relaxes this assumption. Specifically, both the random variable defining an agent's signal and its realization are private information that must be elicited by the designer, increasing the potential for strategic misreporting. This distinction is particularly relevant in practical settings, where assuming common knowledge of each agent's data-generating process—especially in the presence of large and heterogeneous datasets—is unrealistic. Hence, our results accommodate more heterogeneity and offer a higher degree of robustness (Bergemann and Morris, 2005).[6]

Our data-driven transfer scheme also connects with the literature on mechanisms involving contingent payments and public ex-post information (e.g. the work of Hansen (1985) and Riordan and Sappington (1988), and the follow-on literature). Additionally, a growing body of

[6]Several works, including Choi and Kim (1999) and Brooks (2014), consider mechanisms that elicit agents' beliefs about others' types alongside their type reports. Our equilibrium notion allows us to bypass these considerations.

research examines how data on agent types can facilitate efficient outcomes in environments with adverse selection (Braverman and Chassang, 2022; She et al., 2022; Liang and Madsen, 2024) and dynamically evolving market participants (Jiang et al., 2015). Our focus diverges by exploring how the designer can leverage data on a common, payoff-relevant stochastic factor to design transfers that encourage truthful reporting of agents' preferences and information.

Our first application contributes to the literature on sponsored search auctions. The standard model due to Edelman et al. (2007) and Varian (2007) assumes that bidders have a value per click and ad slots have non-increasing click-through rates. Both studies propose essentially the same refinement of pure Nash equilibrium, showing that under this concept, the Generalized-Second Price (GSP) mechanism yields an efficient allocation. A particular feature of the GSP mechanism is that it can be implemented without knowledge of the click-through rates. This is also true for the VCG mechanism, but requires a more careful implementation in which bidders get charged and credited (Varian, 2009; Varian and Harris, 2014).

Milgrom (2010) argues that soliciting a single bid (value per click) can be beneficial—e.g., it can rule out low-revenue equilibria—but also warns that model mis-specification may cause advertisers' values to misalign with observable clicks.

Motivated by this, Dütting et al. (2019) and Dütting et al. (2024) examine standard position auctions with mis-specified bidding languages; and establish a ranking between standard position auction formats in regard to the format's ability to support an efficient equilibrium, when values follow different click-through rates than those used in the auction. Our work approaches this differently, by modeling the click-through rates as private information that the bidders can report to the auctioneer.

Another related line of work is Bergemann et al. (2022) and Chen et al. (2023), who consider a situation where click-through rates are stochastic; and the auctioneer strategically discloses information about the distribution of click-through rates so as to maximize revenue. In this model the auctioneer has additional information, which is in contrast to our model where it is the bidders that hold additional information.

We also contribute to the nascent strand of literature on mechanism design for LLM-based auctions, which began with Dütting et al. (2024), who proposed a sequential auction mechanism for creative ad generation where advertisers bid on a token-by-token basis. Soumalias et al. (2024) propose an auction that lets agents shape the LLM's output via reported reward functions, with costs based on deviations from a reference LLM. Dubey et al. (2024) study bidding for prominence in LLM-generated summaries, and Hajiaghayi et al. (2024) propose a mechanism that probabilistically selects ads in a Retrieval-Augmented Generation framework. In contrast to these approaches, we examine a setting where agents not only have preferences over the central LLM's output but also possess valuable information that can improve its accuracy. Additionally, we explore a richer environment in terms of potential manipulations and mechanism design.

1.5 Outline

The paper is organized as follows. Section 2 introduces the formal model. In Section 3, we analyze the implementation problem in the setting of standard message-driven mechanisms,

establishing an impossibility result for implementation in posterior equilibrium. In Section 4, we define data-driven mechanisms, introduce the data-driven VCG mechanism, and present our implementation results within this framework. Sections 5.1 and 5.2 apply these results to click-through auctions and LLM-based environments, respectively. Finally, Section 6 discusses the implications of our model, summarizes the key findings, proposes directions for future research, and concludes the paper.

2 Model

We consider the following mechanism design setting. There is a set $N \equiv \{1, \ldots, n\}$ of agents. Let X be a compact space of feasible allocations.

Payoffs. Payoff uncertainty is represented by a set of possible states of the world Ω, a compact metric space endowed with the corresponding Borel σ-algebra $\mathcal{B}(\Omega)$. We denote the metric on Ω by d_Ω; we follow this notational convention for all metric spaces considered in this paper. Each agent is endowed with a private *preference type* $\theta_i \in \Theta_i$, where Θ_i is a measurable space. Define $\Theta = \prod_{i \in N} \Theta_i$. Agent i's ex-post utility is quasi-linear in the payoff and transfers: $U_i : X \times \Omega \times \Theta_i \times \mathbb{R} \to \mathbb{R}$ given by:

$$U_i(x, \omega, \theta_i, t_i) = u_i(x, \omega, \theta_i) + t_i,$$

where we assume u_i is common knowledge. We assume u_i is bounded, continuous in x for each θ_i and ω, and measurable in ω for each x and θ_i.

Information. There is a commonly known probability measure $\pi \in \Delta(\Omega)$ with full support. Each agent is endowed with private information about the common payoff-relevant state $\omega \in \Omega$. Drawing on Gentzkow and Kamenica (2017) and Green and Stokey (2022), the underlying stochastic structure is as follows. We extend the state space $(\Omega, \mathcal{B}(\Omega), \pi)$ to the *extended state space* $(\Omega \times [0,1], \mathcal{B}(\Omega) \otimes \mathcal{B}([0,1]), \pi \times \lambda)$. For each agent i, let $\mathcal{S}_i$ be a compact metric space of possible signal realizations endowed with the corresponding Borel σ-algebra $\mathcal{B}(\mathcal{S}_i)$. Define $\mathcal{S} = \prod_{i \in N} \mathcal{S}_i$, with the corresponding product σ-algebra. Each agent i's information is embodied in a measurable function $\mathbf{S}_i : \Omega \times [0,1] \to \mathcal{S}_i$. The mapping $\mathbf{S}_i$ will be referred to as i's *signal* and is assumed to be agent i's private information. Denote by Σ_i the set of all feasible signals for agent i and by $\Sigma = \prod_{i \in N} \Sigma_i$ the product space. We assume Σ is endowed with a metric d_Σ.[7] Each $\mathbf{S}_i \in \Sigma_i$ induces a random variable with the corresponding law given by $P_{\mathbf{S}_i} \equiv (\pi \times \lambda) \circ \mathbf{S}_i^{-1}$. Similarly, a profile of signals $\mathbf{S} \in \Sigma$ induces a random variable with the corresponding joint distribution $P_{\mathbf{S}} \equiv (\pi \times \lambda) \circ \mathbf{S}^{-1}$. Preference types are assumed to be independent of the payoff relevant state. A signal realization $s = \mathbf{S}(\omega, r) \in \mathcal{S}$ leads agents to update their beliefs.[8] The posterior is given by a regular conditional probability distribution given by a Markov kernel

[7] The choice of a specific metric on this space of measurable functions is immaterial for our results.

[8] As we assume preference types are independent of the state, the realized preference types provide no information about the state.

$\pi_{\mathbf{S}} : \mathcal{B}(\Omega) \times \mathcal{S} \to [0,1]$:[9]

$$\pi_{\mathbf{S}}(B|\omega) = \lambda(\{r \in [0,1] : \mathbf{S}(\omega, r) \in B\}) \quad \forall B \in \mathcal{B}(\mathcal{S}), \omega \in \Omega.$$

We define i's expected payoff $v_i : X \times \Theta_i \times \mathcal{S} \times \Sigma \to \mathbb{R}$ as follows:

$$v_i(x, \theta_i, s, \mathbf{S}) = \int_{\Omega} u_i(x, \omega, \theta_i) d\pi_{\mathbf{S}}(\omega|s).$$

Define $\Xi_i \equiv \{(\theta_i, s_i, \mathbf{S}_i) \in \Theta_i \times \mathcal{S}_i \times \Sigma_i : s_i \in \operatorname{supp} P_{\mathbf{S}_i}\}$ to be i's *type space*, with a typical element ξ_i. Define $\Xi = \prod_{i \in N} \Xi_i$ to be the product space. We note that the type space of agents is larger than that considered in the related literature. Specifically, an agent's type includes not only the agent's signal realization but also the random variable governing the information technology itself. Finally, an *instance* is a tuple $\Gamma = (N, X, \Omega, \pi, \Xi, (U_i)_{i \in N})$.

Efficiency. This paper focuses on implementing the efficient allocation rule based on the combined information held by agents while allowing for residual uncertainty about the state of the world.

Definition 1 (Efficient allocation rule)**.** The deterministic allocation rule $x : \Xi \to X$ is efficient if, for all $\xi = (\theta, s, \mathbf{S}) \in \Xi$, it satisfies:[10]

$$x(\xi) \in \arg\max_{x \in X} \sum_{i \in N} v_i(x, \theta_i, s, \mathbf{S}). \tag{1}$$

The argmax correspondence has non-empty and compact values by our continuity and compactness assumptions. We fix an arbitrary selection x^* from the argmax correspondence.

3 Message-Driven Mechanisms

In a message-driven mechanism, agents are asked to select from a menu of messages, and the allocation and transfers are determined solely based on the agents' selection. This is, of course, a standard notion in the mechanism design literature. Our solution concept formally defined below admits a revelation principle by standard arguments. Hence, it is without loss of generality to focus on direct mechanisms and truthtelling. The standard timeline is illustrated in Figure 1.

Definition 2 (Message-driven direct mechanism)**.** A *message-driven direct mechanism* is a pair (x, t) where $x : \Xi \to X$ is the outcome function and $t : \Xi \to \mathbb{R}^N$ is the transfer function.

We focus on efficient implementation in posterior equilibrium, an equilibrium notion first defined by Green and Laffont (1987).

[9] The Markov kernel is $P_{\mathbf{S}}$-a.e. unique. To avoid issues with possible non-uniqueness on measure zero sets, we assume there is a fixed, commonly known, regular conditional probability distribution for each $\mathbf{S} \in \Sigma$. Whenever such an ambiguity arises in later sections, we make an analogous assumption.

[10] We could alternatively formulate the efficient allocation rule as a function of agents' preference types and posterior beliefs. The latter is determined by agents' signals $\mathbf{S}$ and realizations s. We use the current formulation for ease of exposition, as the allocation in subsequent sections will be determined based on agents' reports of their types.

Figure 1: Timing of a message-driven direct revelation mechanism (Definition 3).

Definition 3 (Posterior equilibrium with message-driven mechanisms)**.** A message-driven direct mechanism (x, t) *permits implementation in posterior equilibrium* if for each agent $i \in N$ and types $\xi = (\theta, s, \mathbf{S}) \in \Xi$,

$$v_i(x(\xi), \theta_i, s, \mathbf{S}) + t_i(\xi) \geq v_i(x(\xi_i', \xi_{-i}), \theta_i, s, \mathbf{S}) + t_i(\xi_i', \xi_{-i}) \quad \forall \xi_i' \in \Xi_i.$$

In essence, each agent finds it optimal to report truthfully if other agents do so, ex-post with respect to agent types (though not necessarily the state).[11] This equilibrium concept is appealing because it does not require agents to specify beliefs over others' types and ensures they will not wish to alter their actions after the allocation is finalized.

Our first result establishes that implementation in posterior equilibrium across all instances is impossible with message-driven mechanisms even when we restrict attention to the case of commonly known signals, i.e. a singleton signal Σ. Under this assumption, Maskin (1992), Dasgupta and Maskin (2000), Jehiel and Moldovanu (2001), and Bergemann and Välimäki (2002) demonstrated the impossibility of implementation with interdependent values when signal realizations are either multi-dimensional or single-dimensional without a suitable single-crossing condition.[12] In our setting, even with single-dimensional signal realizations and a single-crossing condition, types remain multi-dimensional. We obtain the following result.

Theorem 1 (Impossibility)**.** *There is an instance where Σ is a singleton, signal realizations of each agent are single-dimensional, and v_i is supermodular in (x, s_j) for each i and j, but for which no message-driven mechanism implements the efficient allocation rule in posterior equilibrium.*

To prove the impossibility result, we provide the following example, which will also serve as our running example later in the paper.

Example 1 (Quadratic loss)**.** There are two agents, and we assume $|\Sigma| = 1$, $X, \Omega, \Theta_1 = \Theta_2, \mathcal{S}_1 = \mathcal{S}_2 \subset \mathbb{R}$ are closed intervals and X is sufficiently wide to make the objects below well-defined.

[11] Note that the posterior equilibrium concept defined by Green and Laffont (1987) for an arbitrary mechanism requires agents' strategies to be optimal against each other's strategies, based on the specific information revealed by the mechanism. In a direct revelation mechanism, this corresponds to revealing the transmitted messages, i.e., the reported types, leading to the notion defined here. The idea that agents have no regret about truthful reporting after uncertainty about others' types is resolved is often referred to as *ex-post equilibrium* (Bergemann and Välimäki, 2002). While these two concepts coincide in our context, they generally differ (Jehiel et al., 2007). Given the importance of resolving uncertainty about the state in our setting, we adopt the terminology of *posterior* equilibrium for discussing implementation.

[12] Each of these works assumes independent signals, but the results extend to cases of arbitrarily correlated signals when considering implementation in posterior equilibrium.

Agent i's payoff is given by

$$u_i(x, \omega, \theta_i) = -(x - \theta_i - \omega)^2.$$

The efficient allocation thus depends on the preferences and the state. The efficient allocation is then given by:[13]

$$x^*(\theta, s) = \frac{\theta_1 + \theta_2}{2} + \mathbb{E}[\omega|s].$$

That is, θ_i can be interpreted as the *bias* agent i seeks to introduce into the allocation. Moreover, each signal realization s_i contributes the estimate of the state ω.

We assume the following condition on the informativeness of signals.[14] There are signal realizations $s_1, s_1' \in \mathcal{S}_1, s_1' \neq s_1$ and $s_2 \in \mathcal{S}_2$ such that

$$\mathbb{E}[\omega|s_1, s_2] \neq \mathbb{E}[\omega|s_1', s_2]. \tag{$*$}$$

We also assume the mapping $x \mapsto \mathbb{E}[\omega|x, s_{-i}]$ is differentiable and increasing for each i and s_{-i}.[15] Under these assumptions, $\frac{\partial v_i(x,\theta_i,s)}{\partial x \partial s_j} \geq 0$ for any $i, j \in N, x \in X, \theta_i \in \Theta_i$, and $s \in \mathcal{S}$.

Proof of Theorem 1. Consider Example 1. We prove that no message-driven mechanism can implement the efficient allocation rule in posterior equilibrium. We proceed by contradiction. Suppose there exist s_1, s_1', s_2 satisfying $(*)$ and a transfer rule t that depends only on agents' reports such that (x^*, t) implements the efficient allocation in posterior equilibrium. Then t must prevent each agent from having a profitable deviation by misreporting either her preference type or signal realization. Using standard results from the literature, we characterize the class of transfer schemes that deter profitable deviations along each dimension of agents' types. The fact that these two classes of payment rules are distinct helps us reach the desired contradiction.

More specifically, suppose the true profile of signal realizations is $s = (s_1, s_2)$ and is fixed to be reported truthfully by both agents. Consider the class of VCG transfers for agent 1:

$$t_1(\theta, s; h_1) = h_1(\theta_2; s) + \mathbb{E}\left[u_2\big(x^*(\theta, s), \omega, \theta_2\big)|s\right] = h_1(\theta_2; s) - \frac{1}{4}(\theta_1 - \theta_2)^2 - \mathrm{Var}[\omega|s],$$

for an arbitrary function $h_1(\theta_2; s)$ of θ_2. We can define the class of VCG transfers for agent 2 analogously. The VCG mechanism implements the efficient outcome in posterior equilibrium. Moreover, any transfer scheme t, such that (x^*, t) permits implementation in posterior equilibrium, must have this form (Green and Laffont, 1977; Holmström, 1979).[16]

Similarly, if preference types θ are fixed to be reported truthfully, and since agent i's expected utility given a profile of signal realizations s is supermodular in (x, s_j) for each j, the sorting conditions of Proposition 4 of Bergemann and Välimäki (2002) are satisfied. Therefore, the

[13] Since Σ is a singleton, the profile of signals is fixed. We drop $\mathbf{S}$ from conditional expectations for the ease of notation.

[14] Note that the condition is satisfied if the signals of all agents jointly determine the state; for example, in the wallet model: $\omega = \sum_{i \in N} s_i$ (Klemperer, 1998). The result thus directly applies to this benchmark case.

[15] These conditions are satisfied, for example, if $\omega \sim N(\mu, \sigma^2)$ with unknown μ and for each i, $s_i = \omega + \epsilon_i$ with independent $\epsilon_i \sim N(0, \sigma_i^2)$ and known variances σ_i^2. To ensure these distributional assumptions match the current framework where all spaces considered are compact, we can suitably truncate these distributions.

[16] Bergemann and Morris (2005) show that with private values, as is assumed here, posterior implementation is equivalent to dominant strategy implementation.

transfer functions t such that (x^*, t) permits implementation in posterior equilibrium must belong to the class of generalized VCG mechanisms with transfers for agent 1 given by:[17]

$$\int_0^{s_1} \frac{\partial}{\partial x} \mathbb{E}\left[u_2\left(x^*(\theta, v_1, s_2), \omega, \theta_2\right) | v_1, s_2\right] \frac{\partial}{\partial v_1} x^*(\theta, v_1, s_2)\, dv_1 = \\ -\left(\theta_1 - \theta_2\right)\left(\mathbb{E}[\omega|s_1, s_2] - \mathbb{E}[\omega|0, s_2]\right).$$

Since $(\theta_1 - \theta_2)\,\mathbb{E}[\omega|0, s_2]$ does not depend on s_1, the generalized VCG payments for agent 1 are of the form

$$t_1(\theta, s; k_1) = k_1(s_2; \theta) - (\theta_1 - \theta_2)\,\mathbb{E}[\omega|s],$$

for an arbitrary function $k_1(s_2; \theta)$ of s_2.

Towards a contradiction, suppose there is a transfer rule t implementing the efficient outcome in posterior equilibrium. Since the transfer function t must ensure there are no profitable deviations along each dimension, it follows that there are functions h_1 and k_1 such that

$$h_1(\theta_2; s) - k_1(s_2; \theta) = \frac{1}{4}(\theta_1 - \theta_2)^2 + \mathrm{Var}[\omega|s] - (\theta_1 - \theta_2)\,\mathbb{E}[\omega|s].$$

Repeating the above for $s_1' \neq s_1$ and s_2, θ:

$$h_1(\theta_2; s_1', s_2) - k_1(s_2; \theta) = \frac{1}{4}(\theta_1 - \theta_2)^2 + \mathrm{Var}[\omega|s_1', s_2] - (\theta_1 - \theta_2)\,\mathbb{E}[\omega|s_1', s_2].$$

Taking their difference, we obtain

$$h_1(\theta_2; s) - h_1(\theta_2; s_1', s_2) = \mathrm{Var}[\omega|s] - \mathrm{Var}[\omega|s_1', s_2] - (\theta_1 - \theta_2)\left(\mathbb{E}[\omega|s] - \mathbb{E}[\omega|s_1', s_2]\right).$$

Since the right-hand side varies with θ_1 by $(*)$, while the left-hand side does not, we arrive at a contradiction, thus completing the proof. □

If a condition analogous to $(*)$ is not satisfied, implementation in posterior equilibrium remains possible. In a model with finite state and signal spaces, McLean and Postlewaite (2015) demonstrate that if agents have *zero informational size* in the terminology of McLean and Postlewaite (2002), that is, the private information held by any single agent is redundant when combined with the joint information of the other agents,[18] their generalized VCG mechanism allows for implementation in posterior equilibrium.[19] Moreover, McLean and Postlewaite (2015) show that if each agent exerts only a small informational effect on the posterior distribution over states, their generalized VCG mechanism is approximately posterior incentive-compatible.[20] In

[17] This is up to the addition of an arbitrary function of agent 2's report of signal and the profile of preference types.

[18] McLean and Postlewaite (2015) refer to this as the *nonexclusive information* condition.

[19] In the generalized VCG (pivot) mechanism proposed by McLean and Postlewaite (2015), agent i pays the externality they impose on the other agents under the assumption that these agents have access to i's information even if i were absent. In the next section, we extend the standard VCG mechanism along similar lines. However, in our pivot mechanism, we account for two types of effects: first, the cost an agent imposes on others by influencing the allocation in their favor, and second, the value they contribute to improving prediction accuracy by sharing information about the state. For the latter, the agent is compensated, as illustrated in Example 2.

[20] The formalization of approximate posterior incentive compatibility by McLean and Postlewaite (2015) differs from our Definition 7. In their framework, a mechanism is defined as *weakly ϵ posterior incentive compatible* if truthful revelation is posterior incentive compatible up to an additive regret of ϵ, with a conditional probability of at least $1 - \epsilon$.

the limit, as the informational size of all agents approaches zero, the mechanism achieves exact implementation in posterior equilibrium.

Typically, however, we expect the signals of all agents to meaningfully influence predictions about the state. Our broader goal is to design mechanisms that are robust to assumptions on the underlying stochastic structure of signals and universally applicable across diverse settings. To this end, we now turn our attention to data-driven mechanisms.

4 Data-Driven Mechanisms

The previous section showed that there are instances where no message-driven mechanism implements the efficient allocation in posterior equilibrium. In this section, we expand the class of mechanisms and demonstrate that this broader class enables at least approximate implementation, as formally defined below. To achieve this, we relax the assumption that both allocations and transfers depend solely on agents' reports, following an approach similar to Mezzetti (2004).[21] Instead, we assume that additional information about the payoff-relevant state ω becomes available after the allocation is determined and that transfers can condition on this information.

In practice, the designer receives additional data about ω. For our motivating examples, ω may capture user preferences, demand, click-through rates, or characteristics of LLM prompts. The platform collects additional data about the state from user interactions with current or past environments, ad clicks, and demand estimation. In settings with AI-generated content, signals like the time users spend on specific prompts or follow-up queries are also valuable.

This data is used to estimate the payoff-relevant state. We formalize the underlying data-generating process as a random variable whose realization is the estimate, thus forming an *estimator* of the state. The estimate is obtained after the final allocation but before determining the final transfers. The timeline is visualized in Figure 2.

Assumption 1 (Estimator)**.** There is a commonly known measurable mapping

$$\widehat{\omega} : \Omega \times [0,1] \to \Omega,$$

which defines a random variable $\widehat{\omega}$ on the extended state space $(\Omega \times [0,1], \mathcal{B}(\Omega) \otimes \mathcal{B}([0,1]), \pi \times \lambda)$, an *estimator* of the state. For any signal profile $\mathbf{S} \in \Sigma$, the random variable $\widehat{\omega}$ is conditionally independent of $\mathbf{S}$ given the payoff-relevant state. The designer receives a realization of $\widehat{\omega}$ after the final allocation and before the final payments.

We denote by $\kappa : \mathcal{B}(\Omega) \times \Omega \to [0,1]$ the corresponding Markov kernel

$$\kappa(B|\omega) = \lambda(\{r \in [0,1] : \widehat{\omega}(\omega, r) \in B\}) \quad \forall B \in \mathcal{B}(\Omega), \omega \in \Omega,$$

representing the distribution of the estimator conditional on the payoff-relevant state.

By the conditional independence assumption and the law of iterated expectations, the expectation of any integrable function $f : \Omega \to \mathbb{R}$ with respect to the posterior distribution

[21] See Subsection 4.2 for a comparison of Mezzetti (2004)'s approach with our framework.

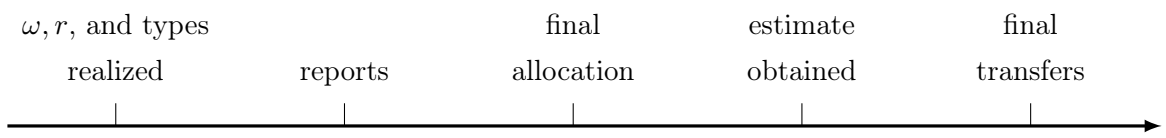

Figure 2: Timing of a data-driven direct revelation mechanism (Definition 4).

$\pi_{\mathbf{S}}(\cdot|s)$—induced by the signal profile $\mathbf{S} \in \Sigma$ and the realization $s \in \mathcal{S}$ (with a slight abuse of notation denoted by $\mathbb{E}[f(\widehat{\omega})|s, \mathbf{S}]$)—is given by

$$\mathbb{E}[f(\widehat{\omega})|s, \mathbf{S}] = \int_{\Omega} \int_{\Omega} f(\widehat{\omega})\, d\kappa(\widehat{\omega}|\omega) d\pi_{\mathbf{S}}(\omega|s) = \int_{\Omega} \int_{[0,1]} f\left(\widehat{\omega}(\omega, r)\right)\, d\lambda(r) d\pi_{\mathbf{S}}(\omega|s).$$

This expression first averages the function f over the randomness in the estimator conditional on the payoff-relevant state, and then takes the expectation with respect to the posterior distribution of the state. The second expression uses that, by construction of the extended state space and the corresponding random variables, the residual randomness embedded in r generates the distribution of the estimator conditional on the payoff-relevant state.

4.1 Mechanisms and Implementation

Data-driven mechanisms condition the allocation only on the reported messages but can condition transfers on an estimate of the payoff-relevant state ω. The formal definition is as follows.

Definition 4 (Data-driven direct mechanism)**.** A *data-driven direct mechanism* is a pair (x, t) where $x : \Xi \to X$ is the outcome function and $t : \Xi \times \Omega \to \mathbb{R}^N$ is the transfer function, where t is integrable with respect to $\kappa(\cdot|\omega)$ for each $\omega \in \Omega$.

The definition of posterior equilibrium (Definition 3) extends naturally to the current setting by incorporating expected transfers, where the expectation is taken with respect to the additional information. Since agents do not observe the realization of the estimator at the reporting stage but know its data-generating process, they form beliefs about its realization. We denote the expected transfer for a profile of signals $\mathbf{S} \in \Sigma$, signal realizations $s \in \mathcal{S}$, and type reports $\xi' \in \Xi$ by

$$\bar{t}_i(\xi', s, \mathbf{S}) \equiv \mathbb{E}[t_i(\xi', \widehat{\omega})|s, \mathbf{S}] = \int_{\Omega} \int_{[0,1]} t_i(\xi', \widehat{\omega}(\omega, r)) d\lambda(r) d\pi_{\mathbf{S}}(\omega|s).$$

Definition 5 (Posterior equilibrium with data-driven mechanisms)**.** A data-driven direct mechanism (x, t) *permits implementation in posterior equilibrium* if for each agent $i \in N$ and types $\xi = (\theta, s, \mathbf{S}) \in \Xi$:

$$v_i(x(\xi), \theta_i, s, \mathbf{S}) + \bar{t}_i(\xi, s, \mathbf{S}) \geq v_i(x(\xi_i', \xi_{-i}), \theta_i, s, \mathbf{S}) + \bar{t}_i(\xi_i', \xi_{-i}, s, \mathbf{S}) \quad \forall \xi_i' \in \Xi_i.$$

With these definitions in place, we define a class of *data-driven VCG* mechanisms by allowing transfers to be conditioned on the realization of the estimator. This class of mechanisms will be the central focus of our analysis in this section.

Definition 6 (Data-driven VCG)**.** A data-driven direct mechanism (x^*, t) is a *data-driven VCG mechanism* if x^* is an efficient allocation rule and for each i, the transfer, as a function of reports $\xi \in \Xi$ and realizations $\widehat{\omega}^*$ of the estimator $\widehat{\omega}$, takes the form

$$t_i(\xi, \widehat{\omega}^*) \equiv h_i(\xi_{-i}, \widehat{\omega}^*) + \sum_{j \neq i} u_j(x^*(\xi), \widehat{\omega}^*, \theta_j),$$

for an arbitrary integrable function h_i of others' reports and realizations $\widehat{\omega}^*$ of the estimator $\widehat{\omega}$.

Any estimator $\widehat{\omega}$ defines a class of data-driven VCG mechanisms. This class is characterized by the specified transfer rule and the requirement that h_i be integrable with respect to the estimator's data-generating process for each agent i. Notably, regardless of the specific choice of the estimator, the class includes all transfer rules defined by functions h_i that are independent of realizations of the estimator for each agent i. Consequently, the intersection of these classes across all feasible estimators is non-empty.

We emphasize a central property of data-driven VCG transfers: they depend on the reported signals only through the resulting allocation. In other words, once the allocation is determined, the designer can compute the transfers without any knowledge of agents' beliefs. We examine various desirable properties of the estimator and explore their implications for implementation.

Ex-Post. As a natural benchmark, we first consider the ex-post case—the designer observes the payoff-relevant state ω after the allocation but before the transfers are finalized. More formally, the estimator fully reveals the state: $\widehat{\omega}(\omega, r) = \omega$ for every $\omega \in \Omega$ and $r \in [0, 1]$.

Theorem 2 (Ex-post)**.** *If $\widehat{\omega}(\omega, r) = \omega$ for every $\omega \in \Omega$ and $r \in [0, 1]$, every data-driven VCG mechanism permits implementation in posterior equilibrium.*

Each agent evaluates data-driven VCG transfers and gross payoffs in expectation using her true beliefs. Thus, if all other agents report truthfully, agent i's net expected payoff aligns with the social objective, ensuring truthfulness by a standard argument.

Proof. Fix an arbitrary data-driven VCG mechanism and agent i. Assume the rest of the agents report truthfully. Then for any $\xi \in \Xi$ and reports $\xi_i' \in \Xi_i$, the expected transfer is given by

$$\bar{t}_i(\xi_i', \xi_{-i}, s, \mathbf{S}) = h_i(\xi_{-i}) + \sum_{j \neq i} v_j(x^*(\xi_i', \xi_{-i}), \theta_j, s, \mathbf{S}),$$

where, with some abuse of notation, we integrated out realizations of the estimator $\widehat{\omega}$ from h_i. The net utility is

$$h_i(\xi_{-i}) + \sum_{j \in N} v_j(x^*(\xi_i', \xi_{-i}), \theta_j, s, \mathbf{S}).$$

Since x^* is efficient, this expression is maximized at ξ_i. □

We illustrate the transfer scheme under the construction of Example 1.

Example 2 (Quadratic loss: data-driven VCG)**.** Example 1 continued. We illustrate that under the pivot version of data-driven VCG transfers, agents are rewarded for providing valuable information about the payoff-relevant state, with rewards increasing in the value of their contribution. However, agents must also pay for steering the social decision in favor of their preferences.

Formally, the pivot version of data-driven VCG transfers is as follows. For any $i, j \in N$ with $i \neq j$,

$$h_i(\theta_j, s_j, \widehat{\omega}^*) \equiv -u_j(x^*(\theta_j, s_j), \widehat{\omega}^*, \theta_j),$$

for any realization $\widehat{\omega}^*$ of $\widehat{\omega}$. For any signal realizations s and reports (θ', s'), the expected transfer is given by:[22]

$$\bar{t}_i(\theta', s', s) = \mathbb{E}\left[-\left(\frac{1}{2}\theta_i' - \frac{1}{2}\theta_j' + \mathbb{E}[\omega|s'] - \omega\right)^2 + \left(\mathbb{E}[\omega|s_j'] - \omega\right)^2 \,\middle|\, s \right].$$

As shown in Proposition 2, there is a posterior equilibrium where all agents report truthfully. In this case, the expected payment of agent i is

$$\bar{t}_i(\theta, s, s) = -\frac{1}{4}(\theta_i - \theta_j)^2 + \left(\mathbb{E}[\omega|s] - \mathbb{E}[\omega|s_j]\right)^2.$$

We highlight the following properties of the payments. First, in expectation, the agent is *paid* for increasing the prediction accuracy. This is most clearly seen if preferences are aligned. Then the transfer of agent i is given by $(\mathbb{E}[\omega|s] - \mathbb{E}[\omega|s_j])^2$. This transfer is always non-negative and strictly positive if agent i's signal induces a different posterior mean when combined with agent j's signal than when based on agent j's signal alone. A necessary condition for this to hold is i having a strictly positive informational size in the terminology of McLean and Postlewaite (2002).

Second, the more accurate the information agent i provides, as measured by a reduction in the expected residual variance $\mathbb{E}[\mathrm{Var}[\omega|s]|s_j] - \mathrm{Var}[\omega|s_j]$, the higher is the agent's expected compensation.[23] Intuitively, $(\mathbb{E}[\omega|s] - \mathbb{E}[\omega|s_j])^2$ measures how much additional information s_i provides about ω beyond s_j; the more informative the signal, the greater the ex-ante expected deviation from the prior mean.

Third, the agent *pays* for introducing bias into the social decision, with the expected payment for introducing bias given by $(\theta_i - \theta_j)^2/4$.

The ex-post case provides a useful benchmark, but the designer likely has only a noisy estimate of the state. Thus, we focus on estimators with two key properties emphasized in statistics and econometrics: unbiasedness and consistency.

Unbiased Estimators. With an unbiased estimator and utility functions affine in the state, data-driven VCG mechanisms achieve implementation in posterior equilibrium. By Jensen's inequality, this result does not extend universally beyond affine utilities.

Proposition 1 (Unbiased estimator)**.** *Suppose $\widehat{\omega}$ is an unbiased estimator of ω conditional on ω pointwise:*

$$\mathbb{E}[\widehat{\omega}|\omega] \equiv \int_{[0,1]} \widehat{\omega}(\omega, r) d\lambda(r) = \omega \quad \forall \omega \in \Omega.$$

[22] For notational ease, since we assume Σ is a singleton, we exclude agents' signals $\mathbf{S}$ from the notation of expected transfers.

[23] Formally, this follows from the law of total variance: $\mathrm{Var}[\mathbb{E}[\omega|s]|s_j] = \mathrm{Var}[\omega|s_j] - \mathbb{E}[\mathrm{Var}[\omega|s]|s_j]$, noting that by the law of iterated expectations, $\mathrm{Var}[\mathbb{E}[\omega|s]|s_j] = \mathbb{E}[(\mathbb{E}[\omega|s] - \mathbb{E}[\omega|s_j])^2|s_j]$. Thus, a reduction in $\mathbb{E}[\mathrm{Var}[\omega|s]|s_j] - \mathrm{Var}[\omega|s_j]$ increases agent i's expected payment.

If utility functions are affine in ω, every data-driven VCG mechanism permits implementation in posterior equilibrium.

Proof. By the law of iterated expectations and conditional independence,

$$\mathbb{E}[\widehat{\omega}|s,\mathbf{S}] = \int_{\Omega}\int_{[0,1]} \widehat{\omega}(\omega,r)d\lambda(r)d\pi_{\mathbf{S}}(\omega|s) = \int_{[0,1]} \omega d\pi_{\mathbf{S}}(\omega|s) = \mathbb{E}[\omega|s,\mathbf{S}].$$

Since utility functions are affine, it follows that for each $i \in N, \mathbf{S} \in \Sigma, s \in \operatorname{supp} P_{\mathbf{S}}, \xi' \in \Xi$:

$$\bar{t}_i(\xi', s, \mathbf{S}) = \sum_{j\neq i} v_j(x^*(\xi'), \theta'_j, s, \mathbf{S}).$$

The result now follows by an argument analogous to the proof of Theorem 2. □

Consistent Estimators. Next, we prove that data-driven VCG mechanisms with a consistent estimator of the state achieve implementation in ϵ-posterior equilibrium (formally defined below), where ϵ can be made arbitrarily small as the estimator converges in probability to the true state.

Definition 7 (ϵ-posterior equilibrium)**.** Fix $\epsilon \geq 0$. A data-driven direct mechanism (x,t) *permits implementation in ϵ-posterior equilibrium* if for each $i \in N$ and types $\xi = (\theta, s, \mathbf{S}) \in \Xi$:

$$v_i(x(\xi),\theta_i,s,\mathbf{S}) + \bar{t}_i(\xi,s,\mathbf{S}) + \epsilon \geq v_i(x(\xi'_i,\xi_{-i}),\theta_i,s,\mathbf{S}) + \bar{t}_i(\xi'_i,\xi_{-i},s,\mathbf{S}) \quad \forall \xi'_i \in \Xi_i.$$

In words, once the uncertainty about others' types is resolved, no agent regrets reporting truthfully by more than ϵ units of the numeraire.

Truthful reporting constitutes an ϵ-posterior equilibrium in finite samples for sufficiently large ϵ regardless of the estimator used. However, with consistent estimators and under the regularity conditions stated below, ϵ can be made arbitrarily small as the size of the dataset used to construct the estimator grows large.

We define the following notions of consistency of estimators. Fixing any $\omega \in \Omega$, we consider the residual randomness generated by r. A sequence of estimators $\{\widehat{\omega}_m\}_m$ is said to be *consistent for* ω if the sequence $\{\widehat{\omega}_m(\omega,\cdot)\}_m$ converges in probability to ω as a random variable on $([0,1],\mathcal{B}([0,1]),\lambda)$:

$$\forall \epsilon > 0: \quad \lim_{m\to\infty} \lambda(\{r \in [0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \epsilon\}) = 0. \tag{2}$$

We define $\{\widehat{\omega}_m\}_m$ to be *pointwise consistent* if it is consistent for every $\omega \in \Omega$:

$$\forall \epsilon > 0: \quad \lim_{m\to\infty} \lambda(\{r \in [0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \epsilon\}) = 0 \quad \forall \omega \in \Omega. \tag{3}$$

Further, $\{\widehat{\omega}_m\}_m$ is *uniformly consistent* if the convergence is uniform:

$$\forall \epsilon > 0: \quad \lim_{m\to\infty} \sup_{\omega\in\Omega} \lambda(\{r \in [0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \epsilon\}) = 0. \tag{4}$$

Before we proceed to state the theorem, we also introduce a continuity notion for posterior beliefs $\pi_{\mathbf{S}}(\cdot|s)$ with respect to signals $\mathbf{S} \in \Sigma$ and signal realizations $s \in \mathcal{S}$. We call posterior beliefs *Lipschitz continuous* if the mapping $(\mathbf{S}, s) \mapsto \pi_{\mathbf{S}}(\cdot|s)$ satisfies the following Lipschitz condition with respect to the Wasserstein 1-distance W_1:

$$\exists L > 0 \text{ s.t. } \forall \mathbf{S}_1, \mathbf{S}_2 \in \Sigma, s_1, s_2 \in \mathcal{S} : W_1(\pi_{\mathbf{S}_1}(\cdot|s_1), \pi_{\mathbf{S}_2}(\cdot|s_2)) \leq L\left(d_{\mathcal{S}}(s_1,s_2) + d_\Sigma(\mathbf{S}_1,\mathbf{S}_2)\right). \tag{5}$$

Theorem 3 (Consistent estimator)**.** *Suppose u_i is Lipschitz in ω uniformly in x and θ_i for each $i \in N$.*[24] *Fix a sequence of estimators $\{\widehat{\omega}_m\}_m$ such that either of the following holds:*

1. *$\{\widehat{\omega}_m\}_m$ is uniformly consistent; or*
2. *$\{\widehat{\omega}_m\}_m$ is pointwise consistent, posterior beliefs are Lipschitz continuous, and Σ is compact.*

Then there is a non-negative sequence $\{\epsilon_m\}_m$, with $\epsilon_m \to 0$ as $m \to \infty$, such that every data-driven VCG mechanism for $\widehat{\omega}_m$ permits implementation in ϵ_m-posterior equilibrium for every $m \in \mathbb{N}$.

Note that when Ω, $\mathcal{S}$, and Σ are finite, pointwise consistency of the estimator is sufficient to obtain the result, as all other regularity conditions hold automatically. In particular, the Lipschitz condition on utility functions follows directly from the finiteness of Ω. Moreover, since Ω is finite, convergence in probability conditional on each $\omega \in \Omega$ in (3) implies uniform convergence in probability across all ω in (4). Likewise, the Lipschitz condition on posteriors in (5) is automatically satisfied when $\mathcal{S}$ and Σ are finite.

The proof proceeds as follows. For any m and types $\xi \in \Xi$, define $\epsilon_i^m(\xi)$ to be the payoff loss for agent i from reporting truthfully compared to i's desired allocation when other agents report truthfully. Also let $\epsilon_m(\xi)$ be the maximum across agents and $\epsilon_m \equiv \sup_{\xi \in \Xi} \epsilon_m(\xi)$. Reporting truthfully is an ϵ_m-posterior equilibrium. Further, as shown in Lemma 1, $\epsilon_m(\xi)$ is upper-bounded by a constant multiple of the difference of expected transfers for the ex-post case and for the m-th estimator $\widehat{\omega}_m$. As shown in Lemma 2, the assumed regularity conditions allow us to further upper-bound this by a suitable supremum over the conditional expectation of the difference between the value of the estimator and the payoff-relevant state. Finally, Lemma 3 shows this upper bound converges to zero.

Fixing an estimator $\widehat{\omega}$, for each agent $i \in N$, with a slight abuse of notation, we define $\bar{t}_i(x, \theta_{-i}, s, \mathbf{S})$ to be the expected transfer under the data-driven VCG mechanism for a profile of signals $\mathbf{S} \in \Sigma$, signal realizations $s \in \mathcal{S}$, allocation $x \in X$, and preference types $\theta_{-i} \in \Theta_{-i}$, omitting the h_i component of data-driven VCG payments:

$$\bar{t}_i(x, \theta_{-i}, s, \mathbf{S}) \equiv \sum_{j \neq i} \mathbb{E}[u_j(x, \widehat{\omega}, \theta_j) | s, \mathbf{S}] = \sum_{j \neq i} \int_{\Omega} \int_{[0,1]} u_j(x, \widehat{\omega}(\omega, r), \theta_j) d\lambda(r) d\pi_{\mathbf{S}}(\omega | s).$$

The first lemma establishes that for any data-driven VCG mechanism for the estimator $\widehat{\omega}$, implementation in an ϵ-posterior equilibrium is feasible for ϵ no larger than a constant multiple of the distance between the expected VCG transfers in the ex-post case and those obtained under $\widehat{\omega}$.

Lemma 1 (Bound on ϵ)**.** *Fix an estimator $\widehat{\omega}$. Then there is an $\epsilon > 0$ with*

$$\epsilon \leq 2 \max_{i \in N} \sup_{\xi = (\theta, s, \mathbf{S}) \in \Xi, x \in X} \left| \sum_{j \neq i} v_j(x, \theta_j, s, \mathbf{S}) - \bar{t}_i(x, \theta_{-i}, s, \mathbf{S}) \right|,$$

[24] That is, $\forall i \in N, \exists L_i > 0$ such that $\forall x \in X, \theta_i \in \Theta_i$ and $\omega_1, \omega_2 \in \Omega$, $|u_i(x, \omega_1, \theta_i) - u_i(x, \omega_2, \theta_i)| \leq L_i d_\Omega(\omega_1, \omega_2)$. Note that a sufficient condition for this to hold is Lipschitz continuity of u_i.

such that every data-driven VCG mechanism for $\widehat{\omega}$ permits implementation in ϵ-posterior equilibrium.

Proof. Fix an estimator $\widehat{\omega}$, a data-driven VCG mechanism for $\widehat{\omega}$, and types $\xi \in \Xi$. Let $i \in N$ be arbitrary. Suppose agents other than i report truthfully. Let $\overline{x}(\xi)$ be a maximizer of i's expected payoff under the estimator for each $\xi = (\theta, s, \mathbf{S}) \in \Xi$:

$$\overline{x}(\xi) \in \arg\max_{x \in X} v_i(x, \theta_i, s, \mathbf{S}) + \overline{t}_i(x, \theta_{-i}, s, \mathbf{S}).$$

This is well-defined by our continuity and compactness assumptions. Define $\epsilon_i(\xi)$ as the utility loss from reporting truthfully:

$$\epsilon_i(\xi) \equiv v_i(\overline{x}(\xi), \theta_i, s, \mathbf{S}) + \overline{t}_i(\overline{x}(\xi), \theta_{-i}, s, \mathbf{S}) - v_i(x^*(\xi), \theta_i, s, \mathbf{S}) - \overline{t}_i(x^*(\xi), \theta_{-i}, s, \mathbf{S}) \geq 0.$$

Let $\epsilon_i \equiv \sup_{\xi \in \Xi} \epsilon_i(\xi)$ and $\epsilon \equiv \max_{i \in N} \epsilon_i$. By construction, truthful reporting constitutes an ϵ-posterior equilibrium. Further, observe that

$$\begin{aligned}\epsilon_i(\xi) = &\sum_{j \in N} v_j(\overline{x}(\xi), \theta_j, s, \mathbf{S}) + \overline{t}_i(\overline{x}(\xi), \theta_{-i}, s, \mathbf{S}) - \sum_{j \neq i} v_j(\overline{x}(\xi), \theta_j, s, \mathbf{S}) \\ &- \sum_{j \in N} v_j(x^*(\xi), \theta_j, s) - \overline{t}_i(x^*(\xi), \theta_{-i}, s, \mathbf{S}) + \sum_{j \neq i} v_j(x^*(\xi), \theta_j, s, \mathbf{S}).\end{aligned}$$

Moreover, since x^* is efficient, $\sum_{j \in N} v_j(\overline{x}(\xi), \theta_j, s, \mathbf{S}) \leq \sum_{j \in N} v_j(x^*(\xi), \theta_j, s, \mathbf{S})$. We obtain an upper bound:

$$\epsilon_i^m = \sup_{\xi \in \Xi} \epsilon_i(\xi) \leq 2 \sup_{\xi = (\theta, s, \mathbf{S}) \in \Xi, x \in X} \left| \sum_{j \neq i} v_j(x, \theta_j, s, \mathbf{S}) - \overline{t}_i(x, \theta_{-i}, s, \mathbf{S}) \right|.$$

Taking a maximum over $i \in N$ on both sides, we obtain the claim. □

Next, we show that the obtained bound can be further bounded from above by the expected distance between the estimator's value and the payoff-relevant state ω while taking the appropriate suprema.

Lemma 2 (Bound via the expected error)**.** *Suppose u_j is Lipschitz in ω uniformly in x and θ_j with a Lipschitz constant L_j for each agent $j \in N$. Then, for each agent $i \in N$,*

$$\sup_{\xi \in \Xi, x \in X} \left| \sum_{j \neq i} v_j(x, \theta_j, s, \mathbf{S}) - \overline{t}_i(x, \theta_{-i}, s, \mathbf{S}) \right| \leq \sum_{j \neq i} L_j \sup_{\mathbf{S} \in \Sigma, s \in \mathcal{S}} \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}(\omega, r), \omega) \, d\lambda(r) \, d\pi_{\mathbf{S}}(\omega | s) \tag{6}$$

$$\leq \sum_{j \neq i} L_j \sup_{\omega \in \Omega} \int_{[0,1]} d_\Omega(\widehat{\omega}(\omega, r), \omega) \, d\lambda(r). \tag{7}$$

The proof uses the assumed Lipschitz condition, the law of iterated expectations, and Jensen's inequality to obtain the result. We include the details in Appendix A.1.

The lemma shows that proving both continuity results hinges on establishing that the upper bounds in (6) and (7) converge to zero. This follows from the maintained regularity and consistency conditions.

Lemma 3 (Uniform convergence)**.** *The following statements hold:*

1. *Suppose $\{\widehat{\omega}_m\}_m$ is a uniformly consistent sequence of estimators. Then,*

$$\lim_{m\to\infty} \sup_{\omega\in\Omega} \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega)\, d\lambda(r) = 0.$$

2. *Suppose $\{\widehat{\omega}_m\}_m$ is a pointwise consistent sequence of estimators, posterior beliefs are Lipschitz continuous, and Σ is compact. Then,*

$$\lim_{m\to\infty} \sup_{\mathbf{S}\in\Sigma, s\in\mathcal{S}} \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega)\, d\lambda(r)\, d\pi_{\mathbf{S}}(\omega|s) = 0.$$

The first part follows by adapting a standard argument for obtaining convergence in expectation using convergence in probability and uniform integrability, which holds in our case by the compactness of Ω, to uniform convergence. The second part is established by showing that the sequence of functions $\{(\mathbf{S}, s) \mapsto \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega)\, d\lambda(r)\, d\pi_{\mathbf{S}}(\omega|s)\}_m$ is uniformly equicontinuous and applying the Arzelà–Ascoli Theorem. We include the details in Appendix A.2.

Combining these results, we obtain the theorem.

Proof of Theorem 3. Fix an arbitrary $m \in \mathbb{N}$ and a data-driven VCG mechanism for $\widehat{\omega}_m$. We obtain an upper bound on feasible ϵ_m such that the data-driven VCG mechanism permits implementation in ϵ_m-posterior equilibrium by applying Lemma 1 and Lemma 2. This upper bound converges to zero by Lemma 3. □

Next, we show that under suitable uniform integrability conditions, the sequence $\{\epsilon_m\}_m$ can converge to zero at essentially the same rate as the sequence of estimators converges to the true state: if $q_m d_\Omega(\widehat{\omega}_m, \omega)$ converges to zero in probability for a non-negative sequence $\{q_m\}_m$, $q_m\epsilon_m$ also converges to zero as $m \to \infty$.[25] Up to a constant factor, this result provides an upper bound on the utility loss agents experience from reporting truthfully when others do the same, given a convergence rate of the estimator.

More formally, we build on (2)-(4) as follows. Fix a non-negative sequence $\{q_m\}_m$. We say a sequence of estimators $\{\widehat{\omega}_m\}_m$ is *consistent for ω at rate $\{q_m\}_m$* if

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \lambda(\{r \in [0,1] : q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) > \epsilon\}) = 0. \tag{8}$$

We define $\{\widehat{\omega}_m\}_m$ to be *pointwise consistent at rate $\{q_m\}_m$* if it is consistent at rate $\{q_m\}_m$ for every $\omega \in \Omega$:

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \lambda(\{r \in [0,1] : q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) > \epsilon\}) = 0 \quad \forall \omega \in \Omega. \tag{9}$$

Further, $\{\widehat{\omega}_m\}_m$ is *uniformly consistent at rate $\{q_m\}_m$* if the convergence is uniform:

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \sup_{\omega\in\Omega} \lambda(\{r \in [0,1] : q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) > \epsilon\}) = 0. \tag{10}$$

[25] That is, $\epsilon_m = o(q_m^{-1})$ in Landau notation.

Unlike Theorem 3, since $\{q_m d_\Omega(\widehat{\omega}_m, \omega)\}_m$ is not guaranteed to be bounded, the following result requires explicit uniform integrability (UI) conditions. We say $\{\widehat{\omega}_m\}_m$ and $\{q_m\}_m$ satisfy *UI pointwise* if

$$\lim_{M\to\infty} \limsup_{m\to\infty} \int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)>M} \lambda(r) = 0 \quad \forall \omega \in \Omega. \tag{11}$$

They satisfy *UI uniformly* if

$$\lim_{M\to\infty} \limsup_{m\to\infty} \sup_{\omega\in\Omega} \int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)>M} \lambda(r) = 0. \tag{12}$$

We are ready to state the formal result.

Proposition 2 (Rate of convergence)**.** *Suppose u_i is Lipschitz in ω uniformly in x and θ_i for each $i \in N$. Fix a sequence of estimators $\{\widehat{\omega}_m\}_m$ and a non-negative sequence $\{q_m\}_m$ such that either of the following holds:*

1. *$\{\widehat{\omega}_m\}_m$ is uniformly consistent at rate $\{q_m\}_m$, and $\{\widehat{\omega}_m\}_m$ and $\{q_m\}_m$ satisfy UI uniformly; or*

2. *$\{\widehat{\omega}_m\}_m$ is pointwise consistent at rate $\{q_m\}_m$, $\{\widehat{\omega}_m\}_m$ and $\{q_m\}_m$ satisfy UI pointwise, posterior beliefs are Lipschitz continuous, and Σ is compact.*

Then there is a non-negative sequence $\{\epsilon_m\}_m$, with $q_m\epsilon_m \to 0$ as $m \to \infty$, such that every data-driven VCG mechanism for $\widehat{\omega}_m$ permits implementation in ϵ_m-posterior equilibrium for every $m \in \mathbb{N}$.

As shown above, by the Lipschitz property of payoff functions, for any profile of signals $\mathbf{S} \in \Sigma$ and signal realizations $s \in \mathcal{S}$, we can upper-bound $q_m\epsilon_m$ by a constant multiple of $\mathbb{E}[q_m d_\Omega(\widehat{\omega}_m, \omega)|s, \mathbf{S}]$. Using the assumed consistency, uniform integrability, and regularity conditions, this converges to 0 uniformly. See Appendix A.3 for details.

4.2 Extensions and Discussion

Eliciting Additional Data from Agents. We have assumed the designer obtains additional information about the state through user engagement, feedback, or third-party sources. Alternatively, suppose that after the allocation is determined, each agent collects additional data about the state, and a second reporting stage is introduced to elicit this information. Based on the reported data, we construct a *leave-one-out* estimator $\widehat{\omega}_{-i}$ for each agent i, derived solely from the information reported by the other agents. The corresponding data-driven VCG transfer is, for any realization $\widehat{\omega}^*_{-i}$ of the estimator $\widehat{\omega}_{-i}$,

$$t_i(\xi, \widehat{\omega}^*_{-i}) \equiv h_i(\xi_{-i}, \widehat{\omega}^*_{-i}) + \sum_{j\neq i} u_j(x^*(\xi), \widehat{\omega}^*_{-i}, \theta_j). \tag{13}$$

Analogously to the two-stage mechanism of Mezzetti (2004), truthful reporting in the second stage is optimal for each agent: since an agent's transfer is independent of the agent's report, agents are indifferent to the reporting choice. Working backward, given truthful reporting in the

second stage, the first stage inherits the incentive properties of the induced games analyzed in this section, provided the corresponding leave-one-out estimators satisfy the required properties. We state this as a result for the ex-post case.

Corollary 1 (Two-stage data-driven VCG in the ex-post case)**.** *Suppose each agent observes the true state after the final allocation but before the final transfers are determined. Consider a two-stage data-driven VCG mechanism: in the first stage, agents report their types, and in the second stage, they report their additional information about the state. Transfers are as in* (13)*. Then there is a perfect Bayesian equilibrium where all agents report truthfully at both stages.*

Alternative Estimators and Information Requirements. We formalized the revelation of additional information through an estimator of the state. An alternative approach would be to base data-driven transfers on estimators of *agent utilities*. With "perfect data," a similar result is obtained if we assume that agents' final utilities, $u_j(x^*(\xi'), \omega, \theta_j)$ for each j, are observed instead; with a slight abuse of notation:

$$t_i(\xi', u) \equiv h_i(\xi'_{-i}, u_{-i}) + \sum_{j \neq i} u_j(x^*(\xi'), \omega, \theta_j).$$

In a model with commonly known signals, i.e. a singleton set of feasible signals Σ, Mezzetti (2004) assumes agents observe their payoffs and report them in a second stage, with transfers given by

$$t_i(\xi', u') \equiv h_i(\xi'_{-i}, u'_{-i}) + \sum_{j \neq i} u'_j,$$

where u' are their reports. There is a perfect Bayesian equilibrium where all agents report truthfully.

Our data-driven mechanisms require *strictly less information* for implementation. Specifically, the final payments do not depend on the knowledge of the agents' true preference types to align incentives. To illustrate, consider the transfer rules in the ex-post case. The "core" of our data-driven VCG transfers for each agent i, based on the reports $\xi' = (\theta', s', \mathbf{S}') \in \Xi$, is given by

$$\sum_{j \neq i} u_j(x^*(\xi'), \omega, \theta'_j),$$

whereas the Mezzetti (2004)-style transfers rely on the knowledge of the agents' *true* payoffs for the chosen allocation. This distinction is further highlighted by comparing our Corollary 1 with Proposition 1 of Mezzetti (2004), which explicitly requires such payoff information.

From a practical standpoint, our data-driven mechanisms also show implementation in posterior equilibrium is possible in cases where a second reporting stage is impractical or prohibitively costly. In such situations, we argue that it is more natural to use the approach developed in this paper. Estimators of agents' payoffs could be derived in the following scenario: after the mechanism is executed and the chosen allocation is implemented, the platform observes user engagement and each agent's prominence—possibly enriched by off-platform data—to estimate agents' payoffs. This requires estimating user engagement, the unobserved component of payoffs. Hence, we argue it is more natural to analyze estimators of the state rather than utilities.

Heterogeneous Priors. While agents might be endowed with heterogeneous and private signals, we have maintained that there is a common prior π over Ω. We can adapt the model to allow for heterogeneous priors as follows. Each agent i is endowed with a full-support prior $\pi_i \in \Pi_i \subseteq \Delta(\Omega)$, with π_i being i's private information. We define an agent-specific extended state space $(\Omega \times [0,1], \mathcal{B}(\Omega) \otimes \mathcal{B}([0,1]), \pi_i \times \lambda)$, which forms agent i's subjective probability space. The rest of the specification follows our baseline model. In particular, a signal of each agent j is a measurable mapping $\mathbf{S}_j : \Omega \times [0,1] \to \mathcal{S}_j$, which is j's private information. Each signal $\mathbf{S}_j$ defines a random variable on i's extended state space, with the law given by $P_{\mathbf{S}_j,\pi_i} = (\pi_i \times \lambda) \circ \mathbf{S}_j^{-1}$. Observing a profile of signals $\mathbf{S}$ and a profile of signal realizations s leads agent i to update her beliefs to a posterior regular conditional distribution $\psi_{\mathbf{S},\pi_i}(\cdot|s)$. We define i's expected payoff as

$$v_i(x, \theta_i, s, \mathbf{S}, \pi_i) = \int_\Omega u_i(x, \omega, \theta_i) d\psi_{\mathbf{S},\pi_i}(\omega|s).$$

Define $\Xi_i \equiv \{(\theta_i, s_i, \mathbf{S}_i, \pi_i) \in \Theta_i \times \mathcal{S}_i \times \Sigma_i \times \Pi_i : s_i \in \operatorname{supp} P_{\mathbf{S}_i,\pi_i}\}$ to be i's type space, and by Ξ the product space across agents. We adjust the definition of efficiency as follows.

Definition 8 (Efficient allocation rule with heterogeneous priors)**.** The deterministic allocation rule $x : \Xi \to X$ is efficient if, for all $\xi = (\theta, s, \mathbf{S}, \pi) \in \Xi$, it satisfies:

$$x(\xi) \in \arg\max_{x \in X} \sum_{i \in N} v_i(x, \theta_i, s, \mathbf{S}, \pi_i). \tag{14}$$

Our implementation results readily extend to this framework and Definition 8. Specifically, we assume there is a commonly known measurable mapping $\widehat{\omega} : \Omega \times [0,1] \to \Omega$, which remains independent of agents' signals conditional on the payoff-relevant state, for any agent i, signal profile $\mathbf{S} \in \Sigma$, and prior $\pi_i \in \Pi_i$. Under this formulation, all properties of estimators discussed in this section remain well-defined across agent types, as we always condition on the payoff-relevant state and treat the estimator as a random variable on the residual probability space $([0,1], \mathcal{B}([0,1]), \lambda)$. Hence, our main results extend naturally under this framework.

Bayesian Interpretation of Additional Information about the State. We defined and analyzed properties of data-driven VCG mechanisms under a primarily frequentist perspective on additional information. A Bayesian approach can also be considered. In Appendix B.1, we formalize additional information about the payoff-relevant state as a further signal independent of agents' signals conditionally on the payoff-relevant state. We identify two ways to define data-driven VCG mechanisms under this framework and analyze their implications for implementation. Here, we provide a brief discussion.

First, we use the posterior mean of the payoff-relevant state given the additional signal as an estimator in data-driven VCG transfers. We obtain analogous results. In particular, we establish that under a suitable version of *posterior consistency* of additional signals, the corresponding sequence of posterior means forms a consistent sequence of estimators (Lemma 4). This, in turn, yields an analogous continuity result to Theorem 3 (Corollary 4).

Second, rather than directly substituting an estimate of the payoff-relevant state into agents' payoffs, we can use the posterior distribution implied by the additional signal to compute expected payoffs and define *Bayesian data-driven VCG* transfers based on them (Definition 11).

Again, we obtain analogous implementation results. In particular, posterior consistency—either uniform or pointwise with a Lipschitz condition on posteriors as in (5)—suffices to establish a continuity result analogous to Theorem 3 (Proposition 3).

State-Revealing Signals. The results in this and the previous section apply to scenarios where agents' signals, when combined, fully reveal the state (see Footnote 14 for a discussion in the context of Theorem 1). This implies that the designer cannot rely solely on reported signals and preference types to determine transfers when agents have strictly positive informational sizes. However, when an estimator with the properties discussed in this section is available, the designer can align incentives. Importantly, even if the estimator is noisy, it cannot simply be replaced by a state constructed directly from agents' reports, as this would allow scope for manipulations.

Multiple Payments. All results presented in this section remain valid if we split the payment into multiple stages, provided that the total payment matches the corresponding transfer schemes. For instance, at the reporting stage, the designer could require a payment equal to the expected VCG payment based on the reported preference type. Once a realization of $\widehat{\omega}$ is observed, the final transfers would then be adjusted to match the corresponding data-driven transfers.

Interdependent Preferences. The arguments presented in this section no longer hold if we introduce additional interdependence into agents' payoffs. Specifically, if the payoff function u_i of each agent i depends on the entire type profile $\theta \in \Theta$, such that $u_i : X \times \Theta \times \Omega \to \mathbb{R}$, the proofs no longer go through. Under these conditions, agent i's report of θ_i influences the expected data-driven VCG transfer not only indirectly, through its effect on the allocation, but also directly, by entering others' payoff functions. The failure is illustrated in Example 5 in Appendix B.

5 Applications

5.1 Click-Through Auctions

In this section, we apply our framework to a canonical model of click-through auctions (Edelman et al., 2007; Varian, 2007). To keep the analysis simple, we consider an auction for a single advertising slot and a fixed number of impressions $K \in \mathbb{N}$.[26] The space of feasible allocations is the probability simplex $X = \Delta_{N-1}$. Each agent is risk-neutral and obtains a fixed payoff $\theta_i \in \mathbb{R}_+$ each time the agent's ad is clicked. We assume a stationary environment and model clicks as binary random variables $Z_{ik} \in \{0, 1\}$ drawn i.i.d. according to the Bernoulli distribution with a parameter $\omega_i \in [0, 1]$—agent i's click-through rate (CTR). Agent i's ex-post

[26]The assumption of a fixed number of impressions is not essential. For example, instead of fixing the number of impressions, the ad display could be auctioned for a specific duration, with impressions arriving stochastically. All objects below can be defined on a per-impression basis, and all results remain valid under this interpretation.

payoff conditional on being allocated the slot is given by

$$\widetilde{u}_i(x, Z_{i1}, \ldots, Z_{iK}, \theta_i) = \theta_i \cdot \sum_{k=1}^{K} Z_{ik}.$$

At the bidding stage, only the expected payoff matters for agents' incentives. Hence, we adopt the following payoff function for each agent i:

$$u_i(x, \omega, \theta_i) = \theta_i \cdot x_i \cdot K \cdot \omega_i,$$

where $K \cdot \omega_i$ represents the expected number of clicks conditional on being allocated the slot. This specification aligns with our framework. The payoff-relevant state is given by CTRs for each agent: $\Omega = [0,1]^N$, with a typical element $\omega = (\omega_1, \ldots, \omega_N)$. Preference types are defined as agents' values per click.

Recall that, in message-driven direct mechanisms, the allocation and transfers depend only on the reported types. This class includes per-impression payments but excludes per-click payments. Following steps similar to Example 1 and assuming the strictly positive informational size condition $(*)$, we can show that no message-driven mechanism implements the efficient allocation in posterior equilibrium in all instances.

Per-click payments fall within the broader class of mechanisms that use data-driven transfers. Under the maintained assumptions, the sample CTR for the displayed ad—the proportion of impressions that result in a click—is an unbiased estimator of the true CTR.[27]

Implementation with per-click pivot (second-price) payments t^{pc}, where for each agent i,

$$t_i^{pc}(\theta, s) = \max_{j \neq i} \theta_j,$$

depends on whether the CTRs are common or individual. First, consider the case of a common CTR, where $\omega = \omega_i = \omega_j$ for each $i, j \in N$. In an efficient allocation, the agent with the highest value per click wins the slot; in the case of ties, any tie-breaker rule is efficient.

Corollary 2 (Common CTR pivot payments)**.** *Suppose there is a common CTR. Then the mechanism (x^*, t^{pc}) with per-click pivot payments t^{pc} permits implementation in posterior equilibrium.*

The expected per-click pivot and data-driven pivot transfers with an estimator given by the sample CTR coincide.[28] Therefore, they must provide the same incentives at the reporting stage.

Proof. Fix an arbitrary agent i and reports ξ'. Agent i expects $K \cdot \mathbb{E}[\omega | s, \mathbf{S}]$ impressions to result in a click. The agent receives the slot with probability $x_i(\xi')$. Thus, the expected transfer is

$$\max_{j \neq i} \theta'_j \cdot x_i^*(\xi') \cdot K \cdot \mathbb{E}[\omega | s, \mathbf{S}].$$

[27] It is also consistent in the limit as $K \to \infty$.

[28] A click on i's ad may not deterministically imply a click on j's ad under a counterfactual slot allocation to j. For example, the set of potential customers might be different. Moreover, clicks of a potential customer across ads might be i.i.d. Nevertheless, through the common CTR, clicks on i's ad allow the designer to construct an unbiased estimator of j's CTR.

Now consider the data-driven pivot transfer $\max_{j\neq i}\theta'_j\cdot x_i^*(\xi')\cdot K\cdot\mathbb{E}[\widehat{\omega}|s,\mathbf{S}]$. By the law of iterated expectations and the fact that the estimator is unbiased, $\mathbb{E}[\widehat{\omega}|s,\mathbf{S}]=\mathbb{E}[\omega|s,\mathbf{S}]$. Therefore, the expected data-driven pivot payment is

$$\max_{j\neq i}\theta'_j\cdot x_i^*(\xi')\cdot K\cdot\mathbb{E}[\omega|s,\mathbf{S}].$$

The result follows from Proposition 1. □

Next, consider the case of individual CTRs. The efficient allocation assigns the slot to agent i if i's expected payoff is the largest among agents, i.e., $\theta_i\cdot\mathbb{E}[\omega_i|s,\mathbf{S}]>\max_{j\neq i}\{\theta_j\cdot\mathbb{E}[\omega_j|s,\mathbf{S}]\}$. With ties, any tie-breaking rule is efficient. However, in this case, the per-click pivot payment does not align incentives. Indeed, fix an arbitrary agent i and assume that all other agents report truthfully. Under the per-click pivot payment, when other agents report truthfully, i's total expected payment is:

$$\max_{j\neq i}\theta_j\cdot x_i^*(\xi'_i,\xi_{-i})\cdot K\cdot\mathbb{E}[\omega_i|s,\mathbf{S}].$$

Observe that the expected payment depends on the reported signals only through the allocation. Consequently, agent i may have an incentive to misreport signals, inflating her own expected CTR while diminishing those of others. By doing so, agent i can secure greater prominence by promising higher payment frequencies, even though these payments are unlikely to materialize based on the agent's posterior beliefs. We illustrate with the following example.

Example 3 (Individual CTR pivot payments)**.** Consider two agents and $K=1$. Suppose that while the prior on ω_i has a non-zero variance, each agent i knows ω_i: Σ_i is a singleton with $\mathcal{S}_i=\Omega_i$ and $\mathbf{S}_i(\omega,r)=\omega_i$ for each $\omega\in\Omega$ and $r\in[0,1]$. Assume further that $\theta_1 s_1<\theta_2 s_2$, but $\theta_1>\theta_2$. If agent 2 reports truthfully and agent 1 reports $\xi'_1=(\theta'_1,s'_1,\mathbf{S}_1)$, the expected net utility of agent 1 under the per-click pivot payment is given by

$$(\theta_1-\theta_2)\cdot x_1^*(\xi'_1,\xi_2)\cdot s_1=(\theta_1-\theta_2)\cdot\mathbb{1}_{\{\theta'_1 s'_1\geq\theta_2 s_2\}}\cdot s_1.$$

Truthfulness is not a posterior equilibrium, as reporting $s'_1,\theta'_1=1$ yields a strictly higher payoff.

With individual CTRs, the corresponding expected data-driven pivot payment for agent i would take the form

$$\bar{t}_i(\xi',s,\mathbf{S})=\max_{j\neq i}\left\{\theta'_j\cdot K\cdot\mathbb{E}[\omega_j|s,\mathbf{S}]\right\}\cdot x_i^*(\xi'). \tag{15}$$

However, implementing such transfers as per-click payments based on agent i's clicks is infeasible, as it would require knowledge of the ratios ω_j/ω_i for each $j\in N\backslash\{i\}$, or an unbiased or consistent estimator of them. If agent i is awarded the slot, the designer would not have such an estimator of other agents' CTRs from the current environment, as other ads are not viewed.[29] Nevertheless, the data-driven pivot mechanism can be implemented if the platform has an estimator of ω_j, potentially sourced from other auction environments or third-party sources.

[29]The same argument holds for any data-driven VCG transfers.

5.2 Mechanism Design for LLM-Generated Contents

We now apply our framework and results to a mechanism design setting where content is generated by an LLM and is desired to be aligned with the private preferences of a set of agents. Such AI alignment problem is extensively studied in LLM training (Ouyang et al., 2022; Rafailov et al., 2024) and has recently attracted much interest from a mechanism design perspective, particularly in settings where agents have conflicting interests (Dütting et al., 2024; Soumalias et al., 2024). LLMs operate as autoregressive generation systems, sequentially predicting the conditional probability of the next token based on the token sequence generated thus far (Brown, 2020).[30] These probabilities are modeled by neural networks trained to maximize the likelihood of the observed text in the training data. For the purpose of our mechanism design framework, we remain agnostic about the specific unit of text output by the LLM—whether it is an individual token or a sequence of tokens forming an output prompt. We collectively refer to this as the "output text." An intrinsic aspect of our analysis is that output texts are randomly generated, which is crucial for the success of LLM technology (Holtzman et al., 2019).

In this context, each input prompt defines a unique mechanism design environment. The output of the LLM is modeled as a unit of text that responds to this prompt. Let T denote the set of feasible outputs, with a typical element $t \in T$. We assume this set is finite. Each agent $i \in N$ is endowed with a *reward function*

$$r_i : T \times \Omega \times \Theta_i \to \mathbb{R},$$

which depends on the generated output text t and the relevant *context* or query-specific information $\omega \in \Omega$. The context captures elements such as user intent or a hidden "ground truth" associated with the prompt. The reward functions are assumed to be be continuous and common knowledge up to a finite-dimensional parameter $\theta_i \in \Theta_i$, which represents the agent's private *preference type*. Each agent i also possesses information about ω in the form of a signal $\mathbf{S}_i$ and a signal realization s_i, which improves the accuracy and relevance of the LLM's output. Signal realizations represent agents' training data or knowledge bases which can help the LLM to generate more desirable content via technologies such as the Retrieval-Augmented Generation (Lewis et al., 2020), while signals represent the data-generating process behind such data. Agents' datasets may overlap, and information across agents may be correlated, as advertisers may share segments of their customer base or source data from the same provider. An important aspect of our model is that it allows for arbitrary correlation among agents' signals. The sets $\Omega, \Theta, \mathcal{S}, \Sigma, \Xi$ follow our earlier specification.

The set of outcomes X corresponds to the set of *generation distributions over output text units*, denoted by $\Delta(T)$. For a given input prompt, architecture, rewards, and data, an LLM produces a generation distribution $x \in X$. Define the *ex-post utility* of agent i from an LLM's generation distribution x as $u_i : X \times \Omega \times \Theta_i \to \mathbb{R}$ such that

$$u_i(x, \omega, \theta_i) = \sum_{t \in T} r_i(t, \omega, \theta_i) \cdot x(t).$$

[30] Tokens, the fundamental units of LLM-generated text, can include subwords, words, phrases, symbols, or numbers. LLMs generate text incrementally on a *token-by-token* basis.

This specification aligns directly with the general framework introduced earlier. We define v_i based on u_i as in Section 2.

Finally, we extend the framework to include a *reference LLM*, x_0, representing the platform's organic generation distribution, similarly to Soumalias et al. (2024). The reference LLM's objective is to provide responses aligned with user queries and maximize their usefulness. However, the platform permits deviations from the reference LLM's output for sufficiently high payments. We measure the magnitude of the deviation of a generation distribution x from the reference generation distribution x_0 with a function

$$\rho : X \times X \to \mathbb{R}_+,$$

mapping the distributions to a non-negative number $\rho(x, x_0)$. Motivated by standard machine learning practice, Soumalias et al. (2024) specify ρ as the Kullback-Leibler divergence

$$D_{KL}(x\|x_0) = \sum_{t\in T} x(t) \log\left(\frac{x(t)}{x_0(t)}\right).$$

We keep the framework general and allows ρ to be any continuous function in x throughout, though will return to this functional form in Example 4.

The designer aims to determine a central LLM generation distribution that maximizes the sum of agents' payoffs while penalizing deviations from the reference LLM distribution.[31] An analogous objective was used by Soumalias et al. (2024); this choice follows standard approaches for tuning LLMs towards human preferences, such as the Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024).[32]

Definition 9 (α-regularized efficient generation distribution)**.** The LLM generation distribution $x : \Xi \to X$ is α-regularized efficient if, for all $\xi = (\theta, s, \mathbf{S}) \in \Xi$, it satisfies

$$x(\xi) \in \arg\max_{x\in X} \mathbb{E}\left[\sum_{t\in T}\sum_{i\in N} r_i(t, \omega, \theta_i)x(t) \;\middle|\; s, \mathbf{S}\right] - \alpha\rho\left(x, x_0\right). \tag{16}$$

Fixing α, we denote by x^* an arbitrary α-regularized efficient generation distribution.

The objective is to implement an α-regularized efficient generation distribution in posterior equilibrium. To achieve this, the platform collects feedback from users' engagement with the central LLM's output prompts and utilizes it to construct an estimator of ω. In the new context of sponsored search auctions via LLMs (Dütting et al., 2024)—where the LLM's output includes advertisers' ads and links—this feedback may consist of not only the observed clicks, as in classic

[31] Another approach would be to specify the platform as an additional agent with a reward function r_0. The results below can be readily adapted to incorporate this framework. Specifically, we could redefine the social objective as a weighted sum of agents' reward functions—including the reward function of the reference LLM agent. This redefinition would remain subject to standard constraints imposed by the LLM architecture. The resulting objective would then guide the specification of a generation distribution that maximizes the constrained social welfare. The data-driven VCG mechanism could be adjusted accordingly.

[32] Unlike in RLHF and DPO, we optimize the sum of many agents' rewards under incentive conflicts, whereas the focus of these papers is optimizing a single agent's reward without considering incentives.

advertising auctions, but also other richer signals. For example, the platform may collect direct user evaluations of answer quality,[33] follow-up queries, and other forms of interaction. These signals collectively inform the estimate of ω. Given this estimate, the modified data-driven VCG transfers are defined as follows.

Definition 10 (α-regularized data-driven VCG)**.** A data-driven direct mechanism (x^*, t) is an *α-regularized data-driven VCG mechanism* if x^* is an α-regularized efficient generation distribution and for each i, the transfer, as a function of reports $\xi \in \Xi$ and realizations $\widehat{\omega}^*$ of the estimator $\widehat{\omega}$, takes the form

$$t_i(\xi, \widehat{\omega}^*) \equiv h_i(\xi_{-i}, \widehat{\omega}^*) + \sum_{j \neq i} u_j(x^*(\xi), \widehat{\omega}^*, \theta_j) - \alpha\rho\left(x^*(\xi), x_0\right),$$

for an arbitrary integrable function h_i of others' reports and realizations $\widehat{\omega}^*$ of the estimator $\widehat{\omega}$.

All results from Section 4 can be readily extended to the current setting. This is formalized in the following corollary, where the statements correspond to Theorem 2, Proposition 1, Theorem 3, and Proposition 2, respectively. For brevity, we omit the proof.

Corollary 3 (Implementation with α-regularized data-driven VCG)**.** *The following holds:*

1. *If $\widehat{\omega}(\omega, r) = \omega$ for every $\omega \in \Omega$ and $r \in [0,1]$, every α-regularized data-driven VCG mechanism permits implementation in posterior equilibrium.*

2. *Suppose $\widehat{\omega}$ is an unbiased estimator of ω conditional on ω pointwise: $\mathbb{E}[\widehat{\omega}|\omega] = \omega$, for every $\omega \in \Omega$. If the reward function of each agent is affine in ω, every α-regularized data-driven VCG mechanism permits implementation in posterior equilibrium.*

3. *Suppose r_i is Lipschitz in ω uniformly in t and θ_i for each $i \in N$. Fix a sequence of estimators $\{\widehat{\omega}_m\}_m$ such that either of the following holds:*

   (a) *$\{\widehat{\omega}_m\}_m$ is uniformly consistent; or*

   (b) *$\{\widehat{\omega}_m\}_m$ is pointwise consistent, posterior beliefs are Lipschitz continuous, and Σ is compact.*

   Then there is a non-negative sequence $\{\epsilon_m\}_m$, with $\epsilon_m \to 0$ as $m \to \infty$, such that every α-regularized data-driven VCG mechanism for $\widehat{\omega}_m$ permits implementation in ϵ_m-posterior equilibrium for every $m \in \mathbb{N}$.

4. *Suppose r_i is Lipschitz in ω uniformly in t and θ_i for each $i \in N$. Fix a sequence of estimators $\{\widehat{\omega}_m\}_m$ and a non-negative sequence $\{q_m\}_m$ such that either of the following holds:*

   (a) *$\{\widehat{\omega}_m\}_m$ is uniformly consistent at rate $\{q_m\}_m$, and $\{\widehat{\omega}_m\}_m$ and $\{q_m\}_m$ satisfy UI uniformly; or*

[33] Several chatbots often prompt users to select from two different outputs, thereby obtaining explicit preference data.

(b) $\{\widehat{\omega}_m\}_m$ *is pointwise consistent at rate* $\{q_m\}_m$*,* $\{\widehat{\omega}_m\}_m$ *and* $\{q_m\}_m$ *satisfy UI pointwise, posterior beliefs are Lipschitz continuous, and* Σ *is compact.*

Then there is a non-negative sequence $\{\epsilon_m\}_m$*, with* $q_m\epsilon_m \to 0$ *as* $m \to \infty$*, such that every* α*-regularized data-driven VCG mechanism for* $\widehat{\omega}_m$ *permits implementation in* ϵ_m*-posterior equilibrium for every* $m \in \mathbb{N}$*.*

We illustrate the construction using D_{KL} as the regularization function.

Example 4 (α-regularized data-driven VCG with KL regularization)**.** Let $\rho = D_{KL}$. Following the derivation by Peters and Schaal (2007) and Rafailov et al. (2024), among others, the α-regularized efficient generation distribution is given by

$$x^*(t|\xi) = \frac{x_0(t)}{Z(\xi)} e^{\frac{1}{\alpha}\sum_{i\in N}\mathbb{E}[r_i(t,\omega,\theta_i)|s,\mathbf{S}]} \quad \text{where} \quad Z(\xi) \equiv \sum_{t\in T} x_0(t)\exp^{\frac{1}{\alpha}\sum_{i\in N}\mathbb{E}[r_i(t,\omega,\theta_i)|s,\mathbf{S}]}.$$

As derived in Appendix B.3, the maximized social objective in (16) is $\alpha\log(Z(\xi))$ and the α-regularized data-driven pivot transfer of agent i is

$$t_i(\xi,\widehat{\omega}^*) = \alpha[\log(Z(\xi)) - \log(Z(\xi_{-i}))] - v_i(x^*(\xi),\theta_i,s,\mathbf{S})$$
$$+\sum_{j\neq i}[u_j(x^*(\xi),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi),\theta_j,s,\mathbf{S})] - [u_j(x^*(\xi_{-i}),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi_{-i}),\theta_j,s_{-i},\mathbf{S}_{-i})].$$

Note that, in the ex-post case where $\widehat{\omega}(\omega,r) = \omega$ for every $\omega \in \Omega$ and $r \in [0,1]$, and assuming a posterior equilibrium in which all agents report truthfully, the expected transfer to agent i simplifies to

$$\bar{t}_i(\xi,s,\mathbf{S}) = \alpha[\log(Z(\xi)) - \log(Z(\xi_{-i})] - v_i(x^*(\xi),\theta_i,s,\mathbf{S}).$$

That is, agent i's expected net payoff is the agent's marginal contribution to the social objective in (16). The intuition from Example 2 extends to this setting as well. Informally, if agents' preference types are sufficiently aligned and agent i's information is valuable for the joint prediction problem, we expect that, ex-ante with respect to signal realizations,

$$\sum_{j\neq i} v_j(x^*(\xi),\theta_j,s,\mathbf{S}) - v_j(x^*(\xi_{-i}),\theta_j,s_{-i},\mathbf{S}_{-i})$$

is positive in expectation. Further, if the difference of the regularization terms

$$D_{KL}(x^*(\xi)\|x_0) - D_{KL}(x^*(\xi_{-i})\|x_0)$$

is relatively small compared to the welfare gain for the other agents, that is, if agent i does not introduce substantial marginal bias into the output of the central LLM, agent i's total transfer is non-negative. We also note that the agent has to pay for introducing bias through two channels: (i) by imposing an allocation externality on the other agents and (ii) by steering the generation distribution away from the reference LLM's distribution.

6 Discussion and Conclusion

We offered an approach to mechanism design that harnesses the natural flow of information in digital environments. By conditioning transfers on post-allocation data, we showed how to achieve implementation even in challenging multi-dimensional settings. Our framework provides a foundation for designing practical mechanisms in modern applications where rich feedback data is readily available.

We also highlighted connections to the literature on efficient mechanism design with interdependent values. In particular, with respect to Mezzetti (2004)'s results, our results provide an insight into the kind and amount of post-allocation data necessary to ensure implementation in at least approximate posterior equilibrium.

Several questions remain open for future research. For example, exploring implementation through message-driven mechanisms in Bayesian equilibrium would be valuable. There is potential for implementing the efficient allocation with correlated types, using constructions similar to the full-surplus-extraction mechanisms of Crémer and McLean (1988) and McAfee and Reny (1992). These could be combined with a version of VCG payments if only one component of the types satisfies Crémer and McLean (1988)'s conditions on the stochastic structure. In such cases, consistent with Neeman (2004)'s insight, agents would retain some information rents. Yet, when these stochastic relevance conditions are not met, it remains unclear whether standard message-driven mechanisms can effectively align incentives. We anticipate that a similar impossibility result for message-driven mechanisms would hold with independent types, following the results of Maskin (1992), Dasgupta and Maskin (2000), and Jehiel and Moldovanu (2001).

Another interesting direction concerns the construction of the estimator. In this paper, we assumed the estimator is exogenous and conditionally independent from agents' information about the state. An important direction for future research is endogenizing the construction of the estimator and relaxing the assumption of conditional independence. For example, it would be interesting to model the estimator as a function of the allocation. In such scenarios, agents might have incentives to misreport to achieve an allocation associated with "weaker" additional information about the state, thereby easing their incentive constraints. This introduces a trade-off between efficiency and incentive alignment: truthful type revelation might require sacrificing efficiency through randomized allocations. Formalizing and analyzing this trade-off further is a fruitful direction.

Building upon the previous points, examining dynamic mechanisms where estimators evolve over multiple iterations within one environment or aggregate information across several environments provides an additional compelling direction. If allocations today influence future estimators, the challenges previously discussed persist. However, dynamic settings could afford the designer richer additional information about the state, as well as a history of reports of an agent, potentially aiding in incentive alignment in Bayesian equilibrium through *linking* environments in the spirit of Jackson and Sonnenschein (2007). Exploring this further is another promising avenue.

Lastly, we assumed a fixed information structure for the agents. From a practical perspective,

incentivizing agents to acquire the socially optimal amount of costly information is an important desideratum, as the central platform aggregates decentralized information from individual agents. Modeling costly information acquisition as an additional stage in the induced game and designing a mechanism that incentivizes the efficient level of information acquisition is another important direction.

References

Bergemann, D., Duetting, P., Paes Leme, R., and Zuo, S. (2022). Calibrated click-through auctions. In *Proceedings of the ACM Web Conference 2022*, pages 47–57.

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73(6):1771–1813.

Bergemann, D. and Välimäki, J. (2002). Information acquisition and efficient mechanism design. *Econometrica*, 70(3):1007–1033.

Braverman, M. and Chassang, S. (2022). Data-driven incentive alignment in capitation schemes. *Journal of Public Economics*, 207:104584.

Brooks, B. A. (2014). Belief extraction in mechanism design.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, J., Li, M., Xu, H., and Zuo, S. (2023). Bayesian calibrated click-through auction. *arXiv preprint arXiv:2306.06554*.

Choi, J. and Kim, T. (1999). A nonparametric, efficient public good decision mechanism: Undominated bayesian implementation. *Games and Economic Behavior*, 27(1):64–85.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public choice*, pages 17–33.

Crémer, J. and McLean, R. P. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pages 1247–1257.

Dasgupta, P. and Maskin, E. (2000). Efficient auctions. *The Quarterly Journal of Economics*, 115(2):341–388.

Dubey, K. A., Feng, Z., Kidambi, R., Mehta, A., and Wang, D. (2024). Auctions with llm summaries. *arXiv preprint arXiv:2404.08126*.

Dütting, P., Fischer, F. A., and Parkes, D. C. (2019). Expressiveness and robustness of first-price position auctions. *Math. Oper. Res.*, 44(1):196–211.

Dütting, P., Fischer, F. A., and Parkes, D. C. (2024). Nontruthful position auctions are more robust to misspecification. *Math. Oper. Res.*, 49(2):901–927.

Dütting, P., Mirrokni, V., Paes Leme, R., Xu, H., and Zuo, S. (2024). Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 144–155.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259.

Folland, G. B. (1999). *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons.

Gentzkow, M. and Kamenica, E. (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429.

Green, J. and Laffont, J.-J. (1977). Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica: Journal of the Econometric Society*, pages 427–438.

Green, J. R. and Laffont, J.-J. (1987). Posterior implementability in a two-person decision problem. *Econometrica: Journal of the Econometric Society*, pages 69–94.

Green, J. R. and Stokey, N. L. (2022). Two representations of information structures and their comparisons. *Decisions in Economics and Finance*, 45(2):541–547.

Groves, T. (1973). Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631.

Hajiaghayi, M., Lahaie, S., Rezaei, K., and Shin, S. (2024). Ad auctions for llms via retrieval augmented generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hansen, R. G. (1985). Auctions with contingent payments. *The American Economic Review*, 75(4):862–865.

Holmström, B. (1979). Groves' scheme on restricted domains. *Econometrica: Journal of the Econometric Society*, pages 1137–1144.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jackson, M. O. and Sonnenschein, H. F. (2007). Overcoming incentive constraints by linking decisions 1. *Econometrica*, 75(1):241–257.

Jehiel, P., Meyer-ter Vehn, M., Moldovanu, B., and Zame, W. R. (2007). Posterior implementation vs ex-post implementation. *Economics Letters*, 97(1):70–73.

Jehiel, P. and Moldovanu, B. (2001). Efficient design with interdependent valuations. *Econometrica*, 69(5):1237–1259.

Jehiel, P. and Moldovanu, B. (2005). Allocative and informational externalities in auctions and related mechanisms. Unpublished manuscript.

Jiang, C., Chen, Y., Wang, Q., and Liu, K. R. (2015). Data-driven auction mechanism design in iaas cloud computing. *IEEE Transactions on Services Computing*, 11(5):743–756.

Klemperer, P. (1998). Auctions with almost common values: Thewallet game'and its applications. *European Economic Review*, 42(3-5):757–769.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Liang, A. and Madsen, E. (2024). Data and incentives. *Theoretical Economics*, 19(1):407–448.

Lu, J. (2019). Bayesian identification: a theory for state-dependent utilities. *American Economic Review*, 109(9):3192–3228.

Maskin, E. (1992). Auctions and privatization. *Privatization*, 1:15–136.

McAfee, R. P. and Reny, P. J. (1992). Correlated information and mecanism design. *Econometrica: Journal of the Econometric Society*, pages 395–421.

McLean, R. and Postlewaite, A. (2002). Informational size and incentive compatibility. *Econometrica*, 70(6):2421–2453.

McLean, R. P. and Postlewaite, A. (2015). Implementation with interdependent valuations. *Theoretical Economics*, 10(3):923–952.

McLean, R. P. and Postlewaite, A. (2017). A dynamic non-direct implementation mechanism for interdependent value problems. *Games and Economic Behavior*, 101:34–48.

Mezzetti, C. (2004). Mechanism design with interdependent valuations: Efficiency. *Econometrica*, 72(5):1617–1626.

Milgrom, P. (2010). Simplified mechanisms with an application to sponsored-search auctions. *Games and Economic Behavior*, 70(1):62–70.

Milgrom, P. R. and Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122.

Neeman, Z. (2004). The relevance of private information in mechanism design. *Journal of Economic theory*, 117(1):55–77.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Peters, J. and Schaal, S. (2007). Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pages 745–750.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Riordan, M. H. and Sappington, D. E. (1988). Optimal contracts with public ex post information. *Journal of Economic Theory*, 45(1):189–199.

She, Z., Ayer, T., and Montanera, D. (2022). Can big data cure risk selection in healthcare capitation program? a game theoretical analysis. *Manufacturing & Service Operations Management*, 24(6):3117–3134.

Soumalias, E., Curry, M. J., and Seuken, S. (2024). Truthful aggregation of LLMs with an application to online advertising. In *Agentic Markets Workshop at ICML 2024*.

Varian, H. R. (2007). Position auctions. *international Journal of industrial Organization*, 25(6):1163–1178.

Varian, H. R. (2009). Online ad auctions. *American Economic Review*, 99(2):430–434.

Varian, H. R. and Harris, C. (2014). The vcg auction in theory and practice. *American Economic Review*, 104(5):442–445.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37.

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Wu, J., Badanidiyuru, A., and Xu, H. (2024). Auctioning with strategically reticent bidders. In *Web and Internet Economics: 20th International Conference, WINE 2024*. Springer.

A Omitted Proofs

A.1 Lemma 2

Proof. Fix an arbitrary agent i. To prove the claim, note that using the Lipschitz property, the law of iterated expectations, and Jensen's inequality, we have $\forall \xi = (\theta, s, \mathbf{S}) \in \Xi$ and $x \in X$:

$$\begin{aligned}\left|\sum_{j\neq i} v_j(x,\theta_j,s,\mathbf{S}) - \bar{t}_i(x,\theta_{-i},s,\mathbf{S})\right| &\leq \sum_{j\neq i}\int_\Omega\int_{[0,1]} |u_j(x,\omega,\theta_j) - u_j(x,\widehat{\omega}(\omega,r),\theta_j)|\, d\lambda(r)d\pi_{\mathbf{S}}(\omega|s)\\ &\leq \sum_{j\neq i} L_j \int_\Omega\int_{[0,1]} d_\Omega(\widehat{\omega}(\omega,r),\omega)d\lambda(r)d\pi_{\mathbf{S}}(\omega|s),\end{aligned}$$

where L_j is the Lipschitz constant on the utility function of agent j.

Furthermore, observe that for any $\mathbf{S} \in \Sigma$ and $s \in \mathcal{S}$, we have

$$\int_\Omega\int_{[0,1]} d_\Omega(\widehat{\omega}(\omega,r),\omega)d\lambda(r)d\pi_{\mathbf{S}}(\omega|s) \leq \sup_{\omega\in\Omega}\int_{[0,1]} d_\Omega(\widehat{\omega}(\omega,r),\omega)d\lambda(r).$$

Hence,

$$\sup_{\mathbf{S}\in\Sigma, s\in\mathcal{S}}\int_\Omega\int_{[0,1]} d_\Omega(\widehat{\omega}(\omega,r),\omega)d\lambda(r)d\pi_{\mathbf{S}}(\omega|s) \leq \sup_{\omega\in\Omega}\int_{[0,1]} d_\Omega(\widehat{\omega}(\omega,r),\omega)d\lambda(r),$$

from which the result follows. □

A.2 Lemma 3

Proof of Part 1. Let $\epsilon > 0$ be arbitrary. Denoting by $\text{diam}(\Omega) = \sup_{\omega,\omega'\in\Omega} d_\Omega(\omega,\omega')$ the diameter of Ω according to the metric d_Ω, we have

$$\begin{aligned}\int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega,r),\omega)d\lambda(r) &= \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{d_\Omega(\widehat{\omega}_m,\omega)\leq\frac{\epsilon}{2}}d\lambda(r) +\\ &\quad \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{d_\Omega(\widehat{\omega}_m,\omega)>\frac{\epsilon}{2}}d\lambda(r)\\ &\leq \frac{\epsilon}{2} + \text{diam}(\Omega)\lambda\left(\left\{r\in[0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \frac{\epsilon}{2}\right\}\right)\end{aligned}$$

Hence,

$$\sup_{\omega\in\Omega}\int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega,r),\omega)d\lambda(r) \leq \frac{\epsilon}{2} + \text{diam}(\Omega)\sup_{\omega\in\Omega}\lambda\left(\left\{r\in[0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \frac{\epsilon}{2}\right\}\right)$$

By the uniform consistency assumption, it follows that there is $M \in \mathbb{N}$ such that, for all $m \geq M$,

$$\sup_{\omega\in\Omega}\lambda\left(\left\{r\in[0,1] : d_\Omega(\widehat{\omega}_m(\omega,r),\omega) > \frac{\epsilon}{2}\right\}\right) < \frac{\epsilon}{2}$$

Hence, for all $m \geq M$:

$$\sup_{\omega\in\Omega}\int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega,r),\omega)d\lambda(r) \leq \epsilon.$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

Proof of Part 2. Define $\Psi_m : \Sigma \times \mathcal{S} \to \mathbb{R}$ by

$$\Psi_m(\mathbf{S}, s) \equiv \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega) d\lambda(r) d\pi_{\mathbf{S}}(\omega|s).$$

Observe that $\Psi_m \to 0$ as $m \to 0$ pointwise. We show this convergence is uniform. To this end, we show the sequence of functions $\{\Psi_m\}_m$ is uniformly equicontinuous and apply the Arzelà–Ascoli Theorem (Folland (1999), Theorem 4.44).

Let $\epsilon > 0$ be arbitrary and fix $\delta = \frac{\epsilon}{\text{diam}(\Omega)L}$. Let $(\mathbf{S}_1, s_1)$ and $(\mathbf{S}_2, s_2)$ be arbitrary elements of $\Sigma \times \mathcal{S}$ such that $d_\Sigma(\mathbf{S}_1, \mathbf{S}_2) + d_\mathcal{S}(s_1, s_2) < \delta$. We have

$$\begin{aligned}
|\Psi_m(\mathbf{S}_1, s_1) - \Psi_m(\mathbf{S}_2, s_2)| &= \left| \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega) d\lambda(r) d[\pi_{\mathbf{S}_1}(\omega|s_1) - \pi_{\mathbf{S}_2}(\omega|s_2)] \right| \\
&\leq \text{diam}(\Omega) \left| \int_\Omega d[\pi_{\mathbf{S}_1}(\omega|s_1) - \pi_{\mathbf{S}_2}(\omega|s_2)] \right| \\
&\leq \text{diam}(\Omega) W_1(\pi_{\mathbf{S}_1}(\cdot|s_1), \pi_{\mathbf{S}_2}(\cdot|s_2)),
\end{aligned}$$

where the last inequality follows by the dual representation of W_1 stemming from the Kantorovich-Rubinstein Theorem (Villani (2021), Theorem 1.14). Using the Lipschitz property (5), we obtain

$$|\Psi_m(\mathbf{S}_1, s_1) - \Psi_m(\mathbf{S}_2, s_2)| \leq \text{diam}(\Omega) L (d_\Sigma(\mathbf{S}_1, \mathbf{S}_2) + d_\mathcal{S}(s_1, s_2)) < \text{diam}(\Omega) L \delta = \epsilon$$

Therefore, $\{\Psi_m\}$ is uniformly equicontinuous. Moreover, the sequence is also uniformly bounded by the compactness of Ω, converges pointwise to a continuous function, and the sequence of functions as well as the pointwise limit is defined on a compact domain by the compactness of $\mathcal{S}$ and Σ. Hence, by the Arzelà–Ascoli Theorem, the convergence is uniform. □

A.3 Proposition 2

Proof. Fix an arbitrary $m \in \mathbb{N}$ and a data-driven VCG mechanism for $\widehat{\omega}_m$. Recall from lemmata 1 and 2 that for each $m \in \mathbb{N}$, there is ϵ_m with

$$\begin{aligned}
\epsilon_m &\leq 2 \max_{i \in N} \sum_{j \neq i} L_j \sup_{\mathbf{S} \in \Sigma, s \in \mathcal{S}} \int_\Omega \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \, d\lambda(r) \, d\pi_{\mathbf{S}}(\omega|s) \\
&\leq 2 \max_{i \in N} \sum_{j \neq i} L_j \sup_{\omega \in \Omega} \int_{[0,1]} d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \, d\lambda(r),
\end{aligned}$$

such that reporting truthfully by all agents is an ϵ_m-posterior equilibrium. It follows that

$$\begin{aligned}
q_m \epsilon_m &\leq 2 \max_{i \in N} \sum_{j \neq i} L_j \sup_{\mathbf{S} \in \Sigma, s \in \mathcal{S}} \int_\Omega \int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \, d\lambda(r) \, d\pi_{\mathbf{S}}(\omega|s) \\
&\leq 2 \max_{i \in N} \sum_{j \neq i} L_j \sup_{\omega \in \Omega} \int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega, r), \omega) \, d\lambda(r).
\end{aligned}$$

We prove the statement for each of the stated conditions. The claim for condition 1 follows by an argument similar to the proof of the first part of Lemma 3. In particular, for any $M > 0$,

we have

$$\sup_{\omega\in\Omega}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\,d\lambda(r) = \sup_{\omega\in\Omega}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\leq M}\,d\lambda(r) + \sup_{\omega\in\Omega}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)>M}\,d\lambda(r).$$

Defining $\mathbf{X}_m(r,\omega) \equiv q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)$, note that for every $\epsilon > 0$ and $\omega \in \Omega$,

$$\lambda(\{r\in[0,1] : \mathbf{X}_m(\omega,r) > \epsilon\}) \leq \lambda(\{r\in[0,1] : q_m d_\Omega(\widehat{\omega}_m(\omega,r) > \epsilon\})$$

By (10), it follows that

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty}\sup_{\omega\in\Omega}\lambda(\{r\in[0,1] : \mathbf{X}_m(\omega,r) > \epsilon\}) = 0.$$

The fact that

$$\forall M > 0 : \quad \lim_{m\to\infty}\sup_{\omega\in\Omega}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\leq M}\,d\lambda(r) = 0$$

follows by an analogous argument to the first part of Lemma 3. Moreover, by the UI condition (12), $\forall\epsilon > 0$, $\exists M > 0$ large enough such that $\forall m \geq M$,

$$\sup_{\omega\in\Omega}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\mathbb{1}_{q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)>M}\,d\lambda(r) < \epsilon$$

Combining the two terms, the claim follows.

To prove the statement for condition 2, note that, by the pointwise convergence (9) and pointwise UI (11) conditions, it follows by the same steps that

$$\lim_{m\to\infty}\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\,d\lambda(r) = 0 \quad \forall\omega\in\Omega.$$

Hence,

$$\lim_{m\to\infty}\int_\Omega\int_{[0,1]} q_m d_\Omega(\widehat{\omega}_m(\omega,r),\omega)\,d\lambda(r)\,d\pi_{\mathbf{S}}(\omega|s) = 0 \quad \forall\mathbf{S}\in\Sigma, s\in\mathcal{S}.$$

The argument showing this convergence is uniform follows analogously to the proof of the second part of Lemma 3.

Hence, in both cases, $\lim_{m\to\infty} q_m\epsilon_m = 0$, concluding the proof. □

B Further Results and Derivations

B.1 Bayesian Interpretation of Additional Data and Alternative Specifications of Data-Driven VCG Mechanisms

In this section, we establish implementation results for an alternative formulation of data-driven VCG mechanisms based on a Bayesian interpretation of additional information about the payoff-relevant state.

Assumption 2 (Additional signal)**.** There is a commonly known compact metric space $\mathcal{Y}$ endowed with the corresponding Borel σ-algebra $\mathcal{B}(\mathcal{Y})$ and a commonly known measurable mapping

$$\mathbf{Y} : \Omega \times [0,1] \to \mathcal{Y},$$

which defines a random variable $\mathbf{Y}$ on the extended state space $(\Omega \times [0,1], \mathcal{B}(\Omega) \otimes \mathcal{B}([0,1]), \pi \times \lambda)$. For any signal profile $\mathbf{S} \in \Sigma$, the random variable $\mathbf{Y}$ is conditionally independent of $\mathbf{S}$ given the payoff-relevant state. The designer receives a realization of $\mathbf{Y}$ after the final allocation and before the final payments.

We denote a generic realization of $\mathbf{Y}$ by $y \in \mathcal{Y}$. We denote by $\kappa_{\mathbf{Y}} : \mathcal{B}(\mathcal{Y}) \times \Omega \to [0,1]$ the corresponding Markov kernel representing the distribution of the signal conditional on the payoff-relevant state. We represent the posterior regular conditional distribution of the payoff-relevant state conditional on the signal using the Markov kernel $\pi_{\mathbf{Y}} : \mathcal{B}(\Omega) \times \mathcal{Y} \to [0,1]$. Finally, we denote the law of $\mathbf{Y}$ by $P_{\mathbf{Y}} = (\pi \times \lambda) \circ \mathbf{Y}^{-1}$.

Interpreting $\mathbf{Y}$ as a signal about the payoff-relevant state, there are at least two ways to define data-driven VCG mechanisms. First, we can define an estimator of the payoff-relevant state as the posterior mean $\widehat{\omega} \equiv \mathbb{E}[\omega|\mathbf{Y}]$. The analysis in this case follows the same steps as in Section 4 using the definition of data-driven VCG mechanisms in Definition 6. In particular, if the signal $\mathbf{Y}$ fully reveals the state, we obtain the ex-post case and exact implementation. A sufficient condition on the signals to obtain the result in Theorem 3, along with the regularity conditions assumed in the theorem, is as follows. Consider a sequence of signals $\{\mathbf{Y}_m\}_m$ that is *posterior consistent* for $\omega \in \Omega$: under the true payoff-relevant state being ω, the sequence of posteriors $\{\pi_{\mathbf{Y}_m}(\cdot|y_m)\}_m$ weakly converges to the Dirac measure on ω, δ_ω, in probability. More formally, since the Wasserstein 1-distance metrizes weak convergence in this setting, we say $\{\mathbf{Y}_m\}_m$ is *posterior consistent for* $\omega \in \Omega$ if

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \kappa_{\mathbf{Y}_m} \left(\{ y : W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) > \epsilon \} \,|\omega \right) = 0. \tag{17}$$

We say $\{\mathbf{Y}_m\}_m$ is *pointwise posterior consistent* if (18) holds for every $\omega \in \Omega$:

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \kappa_{\mathbf{Y}_m} \left(\{ y : W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) > \epsilon \} \,|\omega \right) = 0 \quad \forall \omega \in \Omega. \tag{18}$$

We say $\{\mathbf{Y}_m\}_m$ is *uniformly posterior consistent* if (18) holds uniformly across $\omega \in \Omega$:

$$\forall \epsilon > 0 : \quad \lim_{m\to\infty} \sup_{\omega\in\Omega} \kappa_{\mathbf{Y}_m} \left(\{ y : W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) > \epsilon \} \,|\omega \right) = 0. \tag{19}$$

We obtain the following result, linking pointwise posterior consistency and uniform posterior consistency to pointwise consistency and uniform consistency, respectively, of the estimator of the payoff-relevant state given by the posterior mean.

Lemma 4 (Consistency of the posterior mean)**.** *If $\{\mathbf{Y}_m\}_m$ is pointwise posterior consistent, then the sequence of estimators $\{\widehat{\omega}_m\}_m$ defined by*

$$\widehat{\omega}_m \equiv \mathbb{E}[\omega|\mathbf{Y}_m]$$

for every $m \in \mathbb{N}$ is pointwise consistent. Moreover, if $\{\mathbf{Y}_m\}_m$ is uniformly posterior consistent, the sequence $\{\widehat{\omega}_m\}_m$ is uniformly consistent.

Proof. Suppose $\{\mathbf{Y}_m\}_m$ is posterior consistent for $\omega \in \Omega$. Note that by the definition of the Wasserstein 1-distance, for any signal $\mathbf{Y}$ and $\pi_{\mathbf{Y}}(\cdot|\omega)$-a.e. $y \in \mathcal{Y}$,

$$\begin{aligned} d_\Omega\left(\int \widetilde{\omega} d\pi_{\mathbf{Y}}(\widetilde{\omega}|y), \omega\right) &= d_\Omega\left(\int \widetilde{\omega} d\pi_{\mathbf{Y}}(\widetilde{\omega}|y), \int \widetilde{\omega} d\delta_\omega(\widetilde{\omega})\right) \\ &\leq W_1(\pi_{\mathbf{Y}}(\cdot|y), \delta_\omega). \end{aligned}$$

Fix an arbitrary $\epsilon > 0$. Then, for any $m \in \mathbb{N}$,

$$\lambda(\{r \in [0,1] : d_\Omega(\widehat{\omega}_m(\omega, r), \omega) > \epsilon\}) \leq \kappa_{\mathbf{Y}_m}\left(\{y \in \mathcal{Y} : W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) > \epsilon\}|\omega\right).$$

Taking the limit as $m \to \infty$ yields the first claim. Further, taking the supremum over $\omega \in \Omega$ on both sides and then taking the limit yields the second claim. □

The following then follows directly from Theorem 3.

Corollary 4 (Posterior consistency and the posterior mean)**.** *Suppose u_i is Lipschitz in ω uniformly in x and θ_i for each $i \in N$. Fix a sequence of signals $\{\mathbf{Y}_m\}_m$ such that either (i) $\{\mathbf{Y}_m\}_m$ is uniformly posterior consistent or (ii) $\{\mathbf{Y}_m\}_m$ is pointwise posterior consistent, posterior beliefs are Lipschitz continuous, and Σ is compact. Then there is a non-negative sequence $\{\epsilon_m\}_m$, with $\epsilon_m \to 0$ as $m \to \infty$, such that every data-driven VCG mechanism for $\widehat{\omega}_m \equiv \mathbb{E}[\omega|\mathbf{Y}_m]$ permits implementation in ϵ_m-posterior equilibrium for every $m \in \mathbb{N}$.*

An alternative specification of the transfer rule is to take the expectation of agents' payoffs using the *posterior implied by the signal* $\mathbf{Y}$. The adjusted definition of data-driven VCG mechanisms is as follows.

Definition 11 (Bayesian data-driven VCG)**.** A data-driven direct mechanism (x^*, t) is a *Bayesian data-driven VCG mechanism* if x^* is an efficient allocation rule and for each i,

$$t_i(\xi, y) \equiv h_i(\xi_{-i}, y) + \sum_{j \neq i} \mathbb{E}[u_j(x^*(\xi), \omega, \theta_j)|y, \mathbf{Y}],$$

for an arbitrary integrable function h_i of others' reports and a realization y of the signal $\mathbf{Y}$.

If $\mathbf{Y}$ fully reveals the payoff-relevant state, we again obtain exact posterior implementation following the proof of Theorem 2. Next, we show we also have a continuity result under this formulation, analogously to Theorem 3.

Proposition 3 (Posterior consistency in Bayesian data-driven VCG)**.** *Fix a sequence of signals $\{\mathbf{Y}_m\}_m$ such that either (i) $\{\mathbf{Y}_m\}_m$ is uniformly posterior consistent or (ii) $\{\mathbf{Y}_m\}_m$ is pointwise posterior consistent, posterior beliefs are Lipschitz continuous, and Σ is compact. Then there is a non-negative sequence $\{\epsilon_m\}_m$, with $\epsilon_m \to 0$ as $m \to \infty$, such that every Bayesian data-driven VCG mechanism for $\mathbf{Y}_m$ permits implementation in ϵ_m-posterior equilibrium for every $m \in \mathbb{N}$.*

The proof proceeds similarly to the proof of Theorem 3. In particular, Lemma 1 continues to be a crucial step. We replace Lemma 3 with Lemma 5 stated and proved below. However, we no longer require the Lipschitz condition on payoff functions; hence, we no longer use Lemma

2. Bayesian data-driven VCG mechanisms essentially replace agents' (expected) payoffs with their estimates. Boundedness of payoff function is sufficient to obtain the result by properties of the Wasserstein 1-distance. After proving Lemma 5, we directly proceed to proving the main statements of the proposition.

Lemma 5 (Uniform convergence under posterior consistency)**.** *The following statements hold:*

1. *Fix a uniformly posterior consistent sequence of signals* $\{\mathbf{Y}_m\}_m$*. Then,*

$$\lim_{m\to\infty}\sup_{\omega\in\Omega}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega)=0.$$

2. *Fix a sequence of signals* $\{\mathbf{Y}_m\}_m$ *that is pointwise posterior consistent. Moreover, suppose posterior beliefs are Lipschitz continuous and* Σ *is compact. Then,*

$$\lim_{m\to\infty}\sup_{\mathbf{S}\in\Sigma,s\in\mathcal{S}}\int_{\Omega}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega)d\pi_{\mathbf{S}}(\omega|s)=0.$$

The proof proceeds analogously to the proof of Lemma 3.

Proof. Part 1. Let $\epsilon>0$ be arbitrary. We have

$$\begin{aligned}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega) &= \int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)\mathbb{1}_{W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)\leq\frac{\epsilon}{2}}d\kappa_{\mathbf{Y}_m}(y|\omega)+\\ &\quad\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)\mathbb{1}_{W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)>\frac{\epsilon}{2}}d\kappa_{\mathbf{Y}_m}(y|\omega)\\ &\leq \frac{\epsilon}{2}+\operatorname{diam}(\Omega)\kappa_{\mathbf{Y}_m}\left(\left\{y: W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)>\frac{\epsilon}{2}\right\}|\omega\right)\end{aligned}$$

Hence,

$$\sup_{\omega\in\Omega}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega)\leq\frac{\epsilon}{2}+\operatorname{diam}(\Omega)\sup_{\omega\in\Omega}\kappa_{\mathbf{Y}_m}\left(\left\{y: W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)>\frac{\epsilon}{2}\right\}|\omega\right)$$

By the uniform posterior consistency assumption, it follows that there is $M\in\mathbb{N}$ such that, for all $m\geq M$,

$$\sup_{\omega\in\Omega}\kappa_{\mathbf{Y}_m}\left(\left\{y: W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)>\frac{\epsilon}{2}\right\}|\omega\right)<\frac{\epsilon}{2}$$

Hence, for all $m\geq M$:

$$\sup_{\omega\in\Omega}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega)\leq\epsilon.$$

Since $\epsilon>0$ was arbitrary, the result follows.

Part 2. Define $\Psi_m:\Sigma\times\mathcal{S}\to\mathbb{R}$ by

$$\Psi_m(\mathbf{S},s)\equiv\int_{\Omega}\int_{\mathcal{Y}} W_1(\pi_{\mathbf{Y}_m}(\cdot|y),\delta_\omega)d\kappa_{\mathbf{Y}_m}(y|\omega)d\pi_{\mathbf{S}}(\omega|s).$$

Observe that $\Psi_m\to 0$ as $m\to 0$ pointwise. We show this convergence is uniform. To this end, we show the sequence of functions $\{\Psi_m\}_m$ is uniformly equicontinuous and apply the Arzelà–Ascoli Theorem (Folland (1999), Theorem 4.44).

Let $\epsilon > 0$ be arbitrary and fix $\delta = \frac{\epsilon}{\text{diam}(\Omega)L}$. Let $(\mathbf{S}_1, s_1)$ and $(\mathbf{S}_2, s_2)$ be arbitrary elements of $\Sigma \times \mathcal{S}$ such that $d_\Sigma(\mathbf{S}_1, \mathbf{S}_2) + d_\mathcal{S}(s_1, s_2) < \delta$. We have

$$\begin{aligned}|\Psi_m(\mathbf{S}_1, s_1) - \Psi_m(\mathbf{S}_2, s_2)| &= \left| \int_\Omega \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega) d[\pi_{\mathbf{S}_1}(\omega|s_1) - \pi_{\mathbf{S}_2}(\omega|s_2)] \right| \\ &\leq \text{diam}(\Omega) \left| \int_\Omega d[\pi_{\mathbf{S}_1}(\omega|s_1) - \pi_{\mathbf{S}_2}(\omega|s_2)] \right| \\ &\leq \text{diam}(\Omega) W_1(\pi_{\mathbf{S}_1}(\cdot|s_1), \pi_{\mathbf{S}_2}(\cdot|s_2)),\end{aligned}$$

where the last inequality follows by the dual representation of W_1 stemming from the Kantorovich-Rubinstein Theorem (Villani (2021), Theorem 1.14). Using the Lipschitz property, we obtain

$$|\Psi_m(\mathbf{S}_1, s_1) - \Psi_m(\mathbf{S}_2, s_2)| \leq \text{diam}(\Omega) L (d_\Sigma(\mathbf{S}_1, \mathbf{S}_2) + d_\mathcal{S}(s_1, s_2)) < \text{diam}(\Omega) L \delta = \epsilon$$

Therefore, $\{\Psi_m\}$ is uniformly equicontinuous. Moreover, the sequence is also uniformly bounded by the compactness of Ω, converges pointwise to a continuous function, and the sequence of functions as well as the pointwise limit is defined on a compact domain by the compactness of $\mathcal{S}$ and Σ. Hence, by the Arzelà–Ascoli Theorem, the convergence is uniform. □

Proof of Proposition 3. By Lemma 1, it sufficient to show that for every agent $j \in N$

$$\sup_{\xi=(\theta,s,\mathbf{S}) \in \Xi, x \in X} \left| \int_\Omega u_j(x, \omega, \theta_j) d\pi_\mathbf{S}(\omega|s) - \int_\Omega \int_\mathcal{Y} \int_\Omega u_j(x, \widetilde{\omega}, \theta_j) d\pi_\mathbf{Y}(\widetilde{\omega}|y) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s) \right| \to 0,$$

as $m \to \infty$. To this end, note that for every $m \in \mathbb{N}$, $\xi = (\theta, s, \mathbf{S}) \in \Xi$, and $x \in X$, by Jensen's inequality,

$$\begin{aligned}&\left| \int_\Omega u_j(x, \omega, \theta_j) d\pi_\mathbf{S}(\omega|s) - \int_\Omega \int_\mathcal{Y} \int_\Omega u_j(x, \widetilde{\omega}, \theta_j) d\pi_\mathbf{Y}(\widetilde{\omega}|y) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s) \right| \leq \\ &\qquad \int_\Omega \int_\mathcal{Y} \left| \int_\Omega u_j(x, \widetilde{\omega}, \theta_j) d[\pi_\mathbf{Y}(\widetilde{\omega}|y) - \delta_\omega(\widetilde{\omega})] \right| d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s).\end{aligned}$$

Since we assume throughout this paper that all payoff functions are bounded, and using the Kantorovich-Rubinstein Theorem (Villani (2021), Theorem 1.14), we can further upper-bound this by

$$\|u_j\|_\infty \int_\Omega \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s).$$

Under conditions (ii), by Lemma 5,

$$\begin{aligned}&\sup_{\xi \in \Xi, x \in X} \|u_j\|_\infty \int_\Omega \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s) = \\ &\qquad \|u_j\|_\infty \sup_{\mathbf{S} \in \Sigma, s \in \mathcal{S}} \int_\Omega \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s) \to 0,\end{aligned}$$

as $m \to \infty$. Further, note that

$$\begin{aligned}&\|u_j\|_\infty \sup_{\mathbf{S} \in \Sigma, s \in \mathcal{S}} \int_\Omega \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega) d\pi_\mathbf{S}(\omega|s) \leq \\ &\qquad \|u_j\|_\infty \sup_{\omega \in \Omega} \int_\mathcal{Y} W_1(\pi_{\mathbf{Y}_m}(\cdot|y), \delta_\omega) d\kappa_{\mathbf{Y}_m}(y|\omega).\end{aligned}$$

The right-hand side converges to zero as $m \to \infty$ under condition (i) by Lemma 5. The result follows. □

B.2 Further Examples

Example 5. Suppose there are two agents and a single object to be allocated. The set of allocations is the probability simplex Δ_1. The set of feasible signals Σ is a singleton. The state, preference types, and signal realizations are single-dimensional and all belong to $[0, 1]$. Suppose there are no atoms in the priors on types.

Suppose the payoff of agent 1 is given by

$$u_1(x, \theta, \omega) = \theta_1 \cdot x_1 \cdot \omega,$$

where x_1 is the probability agent 1 obtains the object. The payoff of agent 2 is given by

$$u_2(x, \theta, \omega) = (-\theta_1 + \theta_2) \cdot x_2 \cdot \omega.$$

The efficient allocation is as follows. It is efficient to allocate the object to agent 1 if $2\theta_1 > \theta_2$. If $2\theta_1 < \theta_2$, it is efficient to allocate the object to agent 2. In the case of a tie, any feasible allocation is efficient.

Consider the data-driven VCG expected transfer for the ex-post case:

$$t_1(\theta, s) = h_1(\theta_2, s_2) + \mathbb{E}\left[u_2(x^*(\theta, s), \theta, \omega)|s\right].$$

Suppose agent 2 reports truthfully. Agent 1's expected net utility when reporting (θ_1', s_1') thus becomes

$$h_1(\theta_2, s_2) + \left[\theta_1 x_1^*(\theta_1', \theta_2) + (-\theta_1' + \theta_2) x_2^*(\theta_1', \theta_2)\right] \mathbb{E}\left[\omega|s\right].$$

If $\theta_1 < \theta_2$ and $\mathbb{E}[\omega|s] > 0$, it is optimal for agent 1 to report $\theta_1' = 0$, yielding a payoff of $h_1(\theta_2, s_2) + \theta_2 \mathbb{E}[\omega|s]$. Therefore, reporting truthfully for all agents is not a posterior equilibrium.

B.3 Derivations for Example 4

Recall the maximizer of (16) is

$$x^*(t|\xi) = \frac{x_0(t)}{Z(\xi)} \exp\left(\frac{1}{\alpha} \sum_{i \in N} \mathbb{E}[r_i(t, \omega, \theta_i)|s, \mathbf{S}]\right).$$

To compress notation, define

$$R(t|\xi) \equiv \exp\left(\frac{1}{\alpha} \sum_{i \in N} \mathbb{E}[r_i(t, \omega, \theta_i)|s, \mathbf{S}]\right).$$

Plugging in $x^*(\theta, s)$ into the social objective in 16 yields

$$\begin{aligned}
&\sum_{t \in T} \sum_{i \in N} \mathbb{E}\left[r_i(t, \omega, \theta_i)|s, \mathbf{S}\right] \frac{x_0(t)}{Z(\xi)} R(t|\xi) - \alpha \sum_{t \in T} \frac{x_0(t)}{Z(\xi)} R(t|\xi) \log\left(\frac{R(t|\xi)}{Z(\xi)}\right) \\
&= \sum_{t \in T} \sum_{i \in N} \mathbb{E}\left[r_i(t, \omega, \theta_i)|s, \mathbf{S}\right] \frac{x_0(t)}{Z(\xi)} R(t|\xi) - \sum_{t \in T} \sum_{i \in N} \mathbb{E}\left[r_i(t, \omega, \theta_i)|s, \mathbf{S}\right] \frac{x_0(t)}{Z(\xi)} R(t|\xi) \\
&\quad + \alpha \log\left(Z(\xi)\right) \sum_{t \in T} \frac{x_0(t)}{Z(\xi)} R(t|\xi) \\
&= \alpha \log\left(Z(\xi)\right).
\end{aligned}$$

Now consider

$$\sum_{j\neq i} u_j(x^*(\xi), \widehat{\omega}^*, \theta_j) - \alpha D_{KL}(x^*(\xi)\|x_0),$$

where, as just derived,

$$\begin{aligned}\alpha D_{KL}(x^*(\xi)\|x_0) &= \sum_{t\in T}\sum_{i\in N} \mathbb{E}\left[r_i(t,\omega,\theta_i)|s,\mathbf{S}\right]\frac{x_0(t)}{Z(\xi)}R(t|\xi) - \alpha\log\left(Z(\xi)\right)\\ &= \sum_{i\in N} v_i(x^*(\xi),\theta_i,s,\mathbf{S}) - \alpha\log\left(Z(\xi)\right).\end{aligned}$$

Also note that

$$\sum_{j\neq i} u_j(x^*(\xi), \widehat{\omega}^*, \theta_j) = \sum_{t\in T}\sum_{j\neq i} r_j(t,\widehat{\omega}^*,\theta_j)\frac{x_0(t)}{Z(\xi)}R(t|\xi).$$

Thus,

$$\begin{aligned}&\sum_{j\neq i} u_j(x^*(\xi), \widehat{\omega}^*, \theta_j) - \alpha D_{KL}(x^*(\xi)\|x_0)\\ &= \alpha\log\left(Z(\xi)\right) - v_i(x^*(\xi),\theta_i,s,\mathbf{S}) + \sum_{j\neq i} u_j(x^*(\xi),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi),\theta_j,s,\mathbf{S}).\end{aligned}$$

Similarly,

$$\begin{aligned}&\sum_{j\neq i} u_j(x^*(\xi_{-i}), \widehat{\omega}^*, \theta_j) - \alpha D_{KL}(x^*(\xi_{-i})\|x_0)\\ &= \alpha\log\left(Z(\xi_{-i})\right) + \sum_{j\neq i} u_j(x^*(\xi_{-i}),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi_{-i}),\theta_j,s_{-i},\mathbf{S}_{-i}).\end{aligned}$$

Therefore, the α-regularized data-driven pivot transfer of agent i is

$$\begin{aligned}&\alpha[\log\left(Z(\xi)\right) - \log\left(Z(\xi_{-i})\right] - v_i(x^*(\xi),\theta_i,s,\mathbf{S})\\ &+\sum_{j\neq i}\left[u_j(x^*(\xi),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi),\theta_j,s,\mathbf{S})\right] - \left[u_j(x^*(\xi_{-i}),\widehat{\omega}^*,\theta_j) - v_j(x^*(\xi_{-i}),\theta_j,s_{-i},\mathbf{S}_{-i})\right].\end{aligned}$$