

Agentic AI and the Marginal Utility Computability Dilemma: A Bounded-Measurement Framework

Ahmed Al-Obaid

Saudi Economic Association

May 25, 2026

Abstract

Does agentic AI resolve the marginal utility computability dilemma? This paper addresses this research question by proposing a bounded-measurement framework that formalizes generative agents in preference elicitation. We distinguish between AI-inferred utility, latent utility, and causal identification. Our main theorem demonstrates that although agentic AI expands preference measurement capacity under validation and bounded-error assumptions, it does not directly reveal latent utility and does not independently solve causal identification. Under bounded-error and validation assumptions, calibrated AI utility proxies can improve operational approximation and bound welfare regret in allocation. However, recovering latent utility remains impossible without direct psychological observation, and valid causal inference still requires external identification designs. We conclude that AI-inferred proxies improve measurement and allocation under structured assumptions, but do not reveal latent utility or solve causal identification.

1 Introduction

Since the Marginal Revolution of the late nineteenth century led by William Stanley Jevons, Carl Menger, and Léon Walras, the concept of marginal utility ($MU_i = \partial U / \partial x_i$) has served as the cornerstone of economic value and consumer choice theory. In neoclassical economics, utility represents the latent subjective satisfaction or well-being that an individual derives from consuming a bundle of goods and services. However, from its inception, utility theory has faced a deep structural challenge: utility is a latent, subjective psychological state that cannot be directly observed or cardinally measured (Robbins 1932).

To resolve this measurement dilemma, the discipline underwent several methodological shifts. Hicks and Allen (1934) and Hicks (1939) reconstructed consumer theory on ordinal foundations, showing that consumer choices could be explained entirely by ranking preferences without assigning cardinal values to utility. Samuelson (1938, 1948) pushed this behaviorist transition further with Revealed Preference Theory, arguing that utility is not an unobservable mental state, but is defined operationally by the choices individuals make under budget constraints. Von Neumann and Morgenstern (1944) re-introduced cardinality by analyzing choices under risk (expected utility), showing that preferences over lotteries can be represented by a utility function unique up to a positive affine transformation.

In parallel, the computability of utility and preferences formed the battleground for the **Socialist Calculation Debate** in the 1920s and 1930s. Ludwig von Mises and Friedrich Hayek argued that a central planning authority could never calculate the optimal allocation of resources because the subjective valuations and marginal utilities of millions of individuals are scattered, tacit, and cannot be aggregated. Oskar Lange, by contrast, argued that a central planning board could solve the mathematical market-clearing equations through a trial-and-error process, effectively simulating the market mechanism.

With the recent explosion of Large Language Models (LLMs) and generative agentic AI (such as GPT-4, Claude, and specialized agent architectures), a modern version of this debate has emerged. LLM agents are no longer just passive text generators; they act as active decision-makers, negotiate contracts, choose consumption bundles, and express preferences in experimental environments (Yao et al. 2023; Korinek 2025; Horton 2023). This technological leap raises a fundamental question: **Does agentic AI solve the marginal utility computability dilemma?**

This paper provides a balanced, mathematically rigorous economic review of this question. We show that while agentic AI significantly expands the **operational elicitation** of utility—translating unstructured textual descriptions, behavioral cues, and contexts into precise, quantitative utility proxies—it does not alter the fundamental theoretical limitations of utility theory. The latent nature of utility, the impossibility of objective interpersonal utility comparisons, and the requirement of structural causal design remain intact.

To formalize these insights, we establish a two-layer econometric architecture:

1. **The Measurement Layer**, which maps unstructured textual and behavioral inputs to an LLM-inferred utility proxy, subject to measurement errors and prompt sensitivity.
2. **The Causal Layer**, which embeds this proxy into classical structural models to estimate causal parameters under standard identification strategies (RCT, IV, RD, DiD).

We apply the debiasing framework of Ludwig, Mullainathan, and Rambachan (2025) to show how human validation samples can reduce measurement error and improve construct alignment, while valid causal inference still requires external identification designs such as RCTs, IV, RD, or DiD.

We also address the broader macroeconomic and market implications of agentic AI. We discuss the transition from task-based productivity gains to aggregate Total Factor Productivity (TFP) growth (Acemoglu 2024), and formalize the risk of **knowledge collapse** (Peterson 2024; Acemoglu, Lin, Ozdaglar & Siderius 2026), demonstrating that over-reliance on AI recommendation agents can degrade human learning and shrink the variance of human knowledge. Finally, we analyze the welfare effects of algorithmic pricing, focusing on how personalized first-degree price discrimination extracts consumer surplus and how Q-learning algorithms achieve tacit collusion in repeated Bertrand-Nash games (Calvano et al. 2020; Klein 2021; Fish, Gonczarowski & Shorrer 2024).

The remainder of this paper is structured as follows. Section 2 reviews the historical and theoretical foundations of utility and the socialist calculation debate. Section 3 discusses construct validity and measurement targets. Section 4 presents our mathematical framework, including the random utility model, scale normalization, and causal identification results. Section 5 establishes the falsification conditions for the bounded-computability claim. Section 6 presents applications and extensions, including micro-productivity and macroeconomic growth, the threat of knowledge collapse, algorithmic pricing and collusion, and CES production with agentic capital.

Section 7 provides policy recommendations and concludes.

2 Historical & Theoretical Foundations of Utility

To evaluate the capabilities of agentic AI, we must first place it within the historical development of utility theory. This section details the transitions from ordinalism to expected utility, revealed preference, and the socialist calculation debate.

2.1 Ordinal vs. Cardinal Utility: The Hicks-Allen Framework

Let $X \subset \mathbb{R}_+^n$ be a consumption set representing bundles of n commodities. A consumer's preferences are represented by a binary relation $\succeq$ on X , where $x \succeq y$ denotes that bundle x is at least as preferred as bundle y . Neoclassical consumer theory imposes three core axioms on preferences:

1. **Completeness:** $\forall x, y \in X$, either $x \succeq y$, $y \succeq x$, or both (indifference, written as $x \sim y$).
2. **Transitivity:** $\forall x, y, z \in X$, if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.
3. **Continuity:** For any $y \in X$, the upper contour set $\{x \in X : x \succeq y\}$ and lower contour set $\{x \in X : y \succeq x\}$ are closed in X .

Debreu (1954, 1959) proved that any preference relation satisfying completeness, transitivity, and continuity can be represented by a continuous utility function $U : X \rightarrow \mathbb{R}$ such that:

$$x \succeq y \iff U(x) \geq U(y) \tag{1}$$

Importantly, this representation is **ordinal**. If $g : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing monotonic transformation, then $V(x) = g(U(x))$ represents the exact same preference relation $\succeq$. Under this ordinal framework, the absolute value of utility has no economic meaning, and marginal utility itself is not uniquely defined:

$$\frac{\partial V(x)}{\partial x_i} = g'(U(x)) \frac{\partial U(x)}{\partial x_i} \tag{2}$$

Because $g'(U(x)) > 0$ can take any positive value, the magnitude of marginal utility is arbitrary. However, the Marginal Rate of Substitution (MRS_{ij}), which represents the slope of the

indifference curve, remains invariant to monotonic transformations:

$$MRS_{ij} = \frac{\partial V/\partial x_i}{\partial V/\partial x_j} = \frac{g'(U(x))(\partial U/\partial x_i)}{g'(U(x))(\partial U/\partial x_j)} = \frac{\partial U/\partial x_i}{\partial U/\partial x_j} \quad (3)$$

To visualize this, Figure 1 illustrates the standard ordinal consumer choice framework with indifference curves and budget constraints.

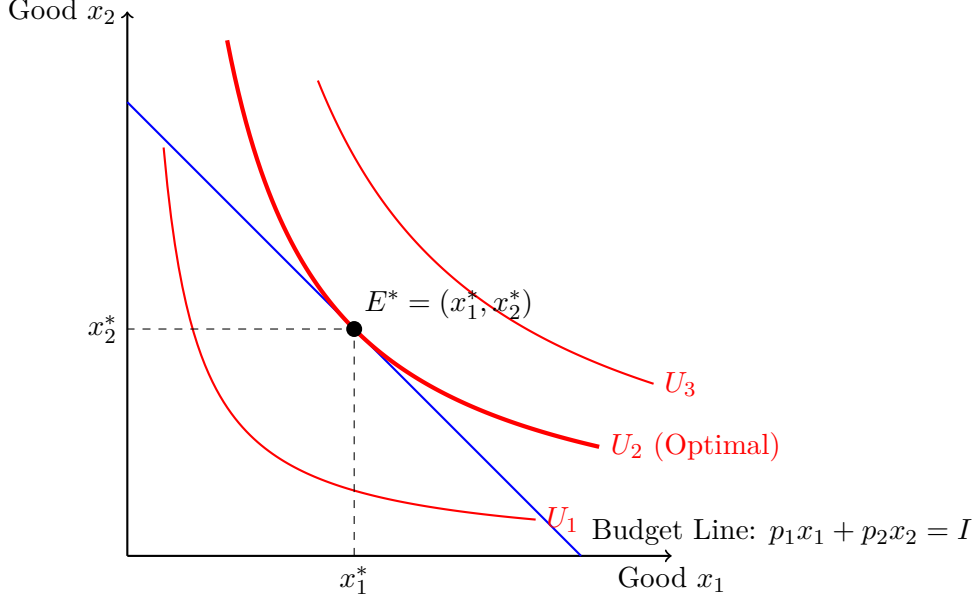


Figure 1: Ordinal Consumer Choice: Indifference Curves and Budget Constraint

Von Neumann and Morgenstern (1944) re-introduced a restricted cardinal representation by extending preferences to the space of lotteries $\mathcal{L}$. If a consumer's preferences over lotteries satisfy the Independence Axiom:

$$L_1 \succeq L_2 \iff \alpha L_1 + (1 - \alpha)L_3 \succeq \alpha L_2 + (1 - \alpha)L_3 \quad \forall \alpha \in (0, 1], L_3 \in \mathcal{L} \quad (4)$$

then there exists a Bernoulli utility function $u : X \rightarrow \mathbb{R}$ such that lottery L_1 is preferred to L_2 if and only if the expected utility under L_1 is higher:

$$L_1 \succeq L_2 \iff \mathbb{E}_{L_1}[u(x)] \geq \mathbb{E}_{L_2}[u(x)] \quad (5)$$

This expected utility representation is unique up to a positive affine transformation $v(x) = au(x) + b$, where $a > 0$. Under this framework, cardinal indicators like the Arrow-Pratt measure of absolute risk aversion ($R_A(x) = -u''(x)/u'(x)$) are uniquely defined (Pratt 1964; Arrow

1965). However, expected utility theory is strictly conditional on choice behavior under risk and does not represent a direct psychological measurement of pleasure or well-being in riskless consumption.

2.2 Revealed Preference and the Afriat Theorem

Samuelson (1938) developed Revealed Preference Theory to bypass the unobservable nature of utility altogether. Let $p^t = (p_t^1, \dots, p_t^n)$ be the price vector at time t , and let $x^t = (x_t^1, \dots, x_t^n)$ be the consumer's chosen bundle. We say that x^t is directly revealed preferred to x , written $x^t \succeq_R x$, if $p^t \cdot x^t \geq p^t \cdot x$ and $x^t \neq x$. The Weak Axiom of Revealed Preference (WARP) states:

$$x^t \succeq_R x \implies x \not\succeq_R x^t \quad \text{i.e., } p^t \cdot x^t \geq p^t \cdot x \text{ and } x^t \neq x \implies p \cdot x < p \cdot x^t \quad (6)$$

To allow transitivity, Houthakker (1950) introduced the Strong Axiom of Revealed Preference (SARP), which applies to chains of choices. Varian (1982) generalized this to the Generalized Axiom of Revealed Preference (GARP):

Definition 1. A dataset $\{(p^t, x^t)\}_{t=1}^T$ satisfies **GARP** if for any sequence $t_1, t_2, \dots, t_k$, if $p^{t_1} \cdot x^{t_1} \geq p^{t_1} \cdot x^{t_2}$, $p^{t_2} \cdot x^{t_2} \geq p^{t_2} \cdot x^{t_3}$, ..., $p^{t_{k-1}} \cdot x^{t_{k-1}} \geq p^{t_{k-1}} \cdot x^{t_k}$, then $p^{t_k} \cdot x^{t_k} \leq p^{t_k} \cdot x^{t_1}$.

Afriat (1967) proved a fundamental theorem showing that GARP is the exact mathematical link between empirical choice data and utility maximization:

Lemma 1 (Afriat's Theorem). For a dataset $\{(p^t, x^t)\}_{t=1}^T$, the following conditions are equivalent:

1. The dataset satisfies GARP.
2. There exists a continuous, concave, and locally non-satiated utility function $U(x)$ that rationalizes the data, such that $x^t \in \arg \max\{U(x) : p^t \cdot x \leq p^t \cdot x^t\}$ for all t .
3. There exist numbers $U^t > 0$ and $\lambda^t > 0$ satisfying the Afriat inequalities:

$$U^s \leq U^t + \lambda^t p^t \cdot (x^s - x^t) \quad \forall s, t \quad (7)$$

Afriat's Theorem implies that we do not need to observe a consumer's internal utility function. As long as their choices satisfy GARP, we can construct a valid utility function mathematically using the Afriat inequalities.

2.3 The Socialist Calculation Debate: Information and Computability

The Socialist Calculation Debate centered on whether a complex economy could be run efficiently without market-determined prices. Ludwig von Mises (1920) argued that rational economic calculation requires a common denominator—monetary prices—which can only arise through the voluntary exchange of private property. Without private ownership of capital goods, there are no markets, no prices, and therefore no way to calculate marginal utility and scarcity relative to capital allocation.

Friedrich Hayek (1935, 1945) expanded this argument by focusing on the nature of information. He argued that the “data” required for central planning do not exist in a single mind or database. Instead, economic knowledge is scattered, tacit, local, and constantly changing:

“The economic problem of society is... a problem of the utilization of knowledge which is not given to anyone in its totality.” (Hayek 1945)

Prices act as a decentralized communication system, aggregating this scattered information. An increase in the price of a raw material signals its scarcity to all users without any central authority having to calculate the individual marginal utilities.

Oskar Lange (1936) counter-argued that a central planning board could act as a market simulator. By setting arbitrary starting prices, observing surpluses and shortages, and adjusting prices according to Walrasian trial-and-error (*tâtonnement*), the planning board could solve the system of equations and achieve Pareto efficiency.

The rise of agentic AI has reopened this debate. If LLM agents can process massive, unstructured datasets and simulate human behavior, could a supercomputer run an economy by estimating and maximizing utility directly? As we show in the following sections, agentic AI cannot bypass the fundamental constraints identified by Mises and Hayek:

- **The Lucas Critique and Data Limits:** LLMs are trained on historical data. They cannot predict how human preferences and marginal utility will evolve when faced with novel situations or institutional changes.
- **Tacit Knowledge:** Subjective marginal utility is not fully captured in text. Much of human preference is revealed only at the moment of choice under real-world scarcity.
- **Control and Incentives:** Centralizing AI utility estimation requires a political choice of what target function to maximize, exposing the system to Arrow’s Impossibility Theorem

(Arrow 1951) and public choice critiques.

3 Construct Validity and Measurement Targets

In empirical economics, a fundamental requirement for any measurement methodology is *construct validity*—the degree to which an operational measure accurately reflects the theoretical construct it is intended to capture. When integrating agentic AI into economic analysis, researchers must distinguish between several distinct concepts of preference, utility, and welfare.

We formalize these distinctions in Table 1. Crucially, AI-inferred utility ($\hat{u}_i^{\text{AI}}$) is a distinct mathematical and operational object. It is not identical to ordinal preference, latent utility, willingness-to-pay (WTP), von Neumann-Morgenstern (vNM) expected utility, social welfare, or consumer surplus. Equivalence can only be asserted under restrictive validation and normalization assumptions.

Without these assumptions, $\hat{u}_i^{\text{AI}}$ is merely a numerical representation of text patterns, browsing history, or behavioral context processed through a specific set of model weights and prompt instructions. For example, to map AI-inferred utility to willingness-to-pay, the researcher must assume a constant and known marginal utility of income to scale the AI score into monetary units. Similarly, interpersonal comparisons of AI-inferred utility across different individuals assume that the AI’s internal scale is calibrated identically for all agents. Researchers must avoid the conflation of these distinct constructs, as treating $\hat{u}_i^{\text{AI}}$ as a direct, unmediated measure of subjective human well-being violates foundational principles of measurement theory.

4 Mathematical Framework and Identification Results

4.1 Basic Setup and the Inferred Utility Signal

To analyze the role of agentic AI in modern economics, we propose a formal framework that separates utility estimation into two distinct layers: the **Measurement Layer** and the **Causal Layer**. This architecture is represented in Figure 2.

Let $i \in \{1, \dots, N\}$ index individuals, $j \in \{1, \dots, J\}$ index alternatives, and $t \in \{1, \dots, T\}$ index time. The latent economic object of interest is the unobserved utility u_{ijt}^* , representing the latent utility that individual i obtains from alternative j at time t .

This utility is not directly observed. Instead, the researcher observes prices, product char-

Table 1: Economic Preference Constructs vs. AI-Inferred Utility

Construct	Economic Definition	Measurement Basis	Relation to AI Proxy
Ordinal Preference ($\succeq$)	An axiomatic binary ordering over choice bundles satisfying completeness, transitivity, and continuity.	Pairwise choice rankings, behavior under constraint.	Not equivalent. AI proxy imposes a cardinal scale and depends on prompt text rather than choice behavior.
Latent Utility (u^*)	A subjective cardinal or ordinal representation of satisfaction derived from consumption (unobservable).	Not directly measurable; reconstructed from choices.	Not equivalent. AI-inferred utility is an operational proxy containing measurement error.
AI-Inferred Utility ($\hat{u}_i^{\text{AI}}$)	A numeric score extracted by an LLM processing text, context, or behavior.	Language model processing of search history, reviews, or prompts.	This is the AI proxy itself, not an economic primitive. It is a model-generated measurement object subject to prompt sensitivity, training-data bias, and calibration error.
Willingness-to-Pay (WTP)	The maximum monetary amount a consumer is willing to give up to acquire a good.	Observed demand elasticity, pricing experiments, auctions.	Not equivalent. Requires scaling by the marginal utility of income, which AI does not directly capture.
vNM Expected Utility	A cardinal representation of preferences over risky lotteries, unique up to positive affine transformation.	Choices under uncertainty; lottery choices.	Not equivalent. AI proxy does not guarantee satisfaction of the independence axiom or risk-neutral scale.
Social Welfare (W)	An aggregation of individual utilities or well-being across society.	Normative social welfare functions, interpersonal weighting.	Not equivalent. Aggregation is constrained by Arrow's Theorem (Arrow 1951); AI cannot solve interpersonal weighting.
Consumer Surplus (CS)	The difference between a consumer's total willingness to pay and the amount paid.	Area under the Marshallian or Hicksian demand curve.	Not equivalent. Requires integration of behavioral demand curves, not just individual text proxies.

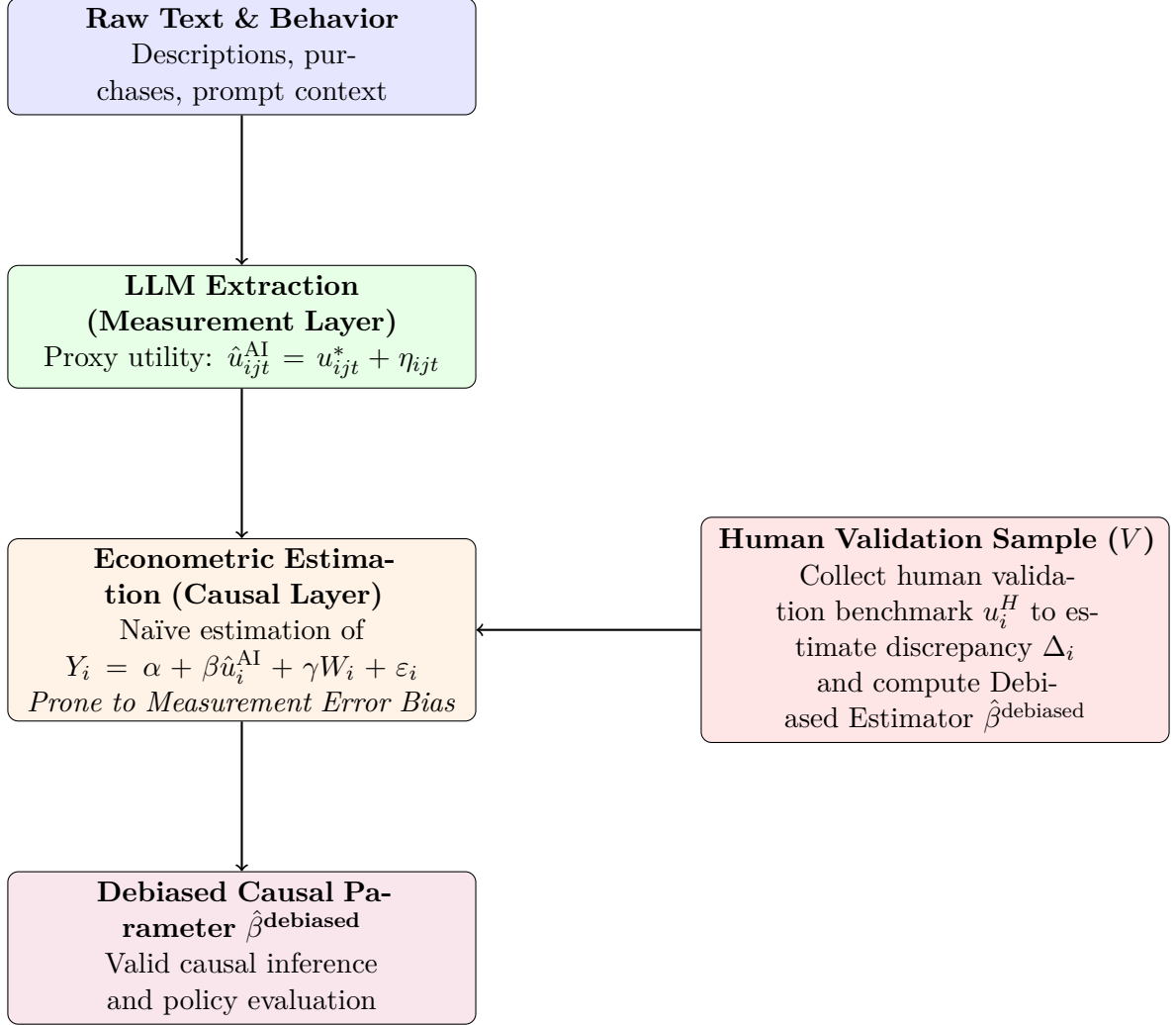


Figure 2: The Two-Layer Econometric Architecture of AI Utility Estimation

acteristics, behavioral traces, and text-based signals:

$$X_{ijt}, \quad p_{jt}, \quad B_{it}, \quad T_{it} \quad (8)$$

where X_{ijt} is a vector of observable characteristics, p_{jt} is price, B_{it} is observed behavior, and T_{it} is unstructured text such as reviews, preference descriptions, search queries, complaints, or conversational data.

The structural latent utility can be written as:

$$u_{ijt}^* = X_{ijt}'\beta - \lambda p_{jt} + \alpha_i + \xi_{ijt} \quad (9)$$

where β captures the valuation of observable characteristics, $\lambda > 0$ is price sensitivity, α_i is individual heterogeneity, and ξ_{ijt} is an unobserved utility component.

The agentic AI system processes these observables and produces an inferred utility signal:

$$\hat{u}_{ijt}^{\text{AI}} = f_{\theta}(T_{it}, B_{it}, X_{ijt}, p_{jt}) \quad (10)$$

where $f_{\theta}(\cdot)$ is the AI extraction function. The key assumption is not that AI observes utility perfectly, but that it produces a noisy proxy:

$$\hat{u}_{ijt}^{\text{AI}} = u_{ijt}^* + \eta_{ijt} \quad (11)$$

where η_{ijt} is measurement error.

This is the central distinction of the paper:

$$\hat{u}_{ijt}^{\text{AI}} \neq u_{ijt}^* \quad (12)$$

in general. AI-inferred utility is a measured proxy, not latent utility itself. Economists must not assume that the AI has directly measured latent utility; rather, it produces an operational proxy whose deviations must be accounted for.

4.2 Information Expansion and Preference Ordinality

Proposition 1. *AI Expands Measurement If It Contains Incremental Information.*

Proof. Let the traditional information set be:

$$\mathcal{I}_0 = \sigma(X_{ijt}, p_{jt}) \quad (13)$$

and the AI-augmented information set be:

$$\mathcal{I}_1 = \sigma(X_{ijt}, p_{jt}, \hat{u}_{ijt}^{\text{AI}}) \quad (14)$$

with:

$$\mathcal{I}_0 \subseteq \mathcal{I}_1 \quad (15)$$

The optimal mean-square predictor of latent utility using information set $\mathcal{I}_k$ is:

$$m_k = \mathbb{E}[u_{ijt}^* \mid \mathcal{I}_k] \quad (16)$$

for $k = 0, 1$. Then:

$$\mathbb{E}[(u_{ijt}^* - m_1)^2] \leq \mathbb{E}[(u_{ijt}^* - m_0)^2] \quad (17)$$

with strict inequality if:

$$\mathbb{E}[u_{ijt}^* \mid \mathcal{I}_1] \neq \mathbb{E}[u_{ijt}^* \mid \mathcal{I}_0] \quad (18)$$

Since $\mathcal{I}_0 \subseteq \mathcal{I}_1$, the set of predictors measurable with respect to $\mathcal{I}_0$ is a subset of the predictors measurable with respect to $\mathcal{I}_1$. The conditional expectation $\mathbb{E}[u_{ijt}^* \mid \mathcal{I}_k]$ is the unique mean-square-error-minimizing predictor of u_{ijt}^* given $\mathcal{I}_k$. Therefore, adding $\hat{u}_{ijt}^{\text{AI}}$ to the information set cannot increase the minimum achievable prediction error.

Thus:

$$\mathbb{E}[(u_{ijt}^* - \mathbb{E}[u_{ijt}^* \mid \mathcal{I}_1])^2] \leq \mathbb{E}[(u_{ijt}^* - \mathbb{E}[u_{ijt}^* \mid \mathcal{I}_0])^2] \quad (19)$$

Strict inequality holds whenever the AI signal contains information about u_{ijt}^* not already contained in X_{ijt} and p_{jt} . $\square$

Interpretation: This proves the first major claim of the paper: AI adoption leads to expanded measurement capability ($\text{AI} \Rightarrow \text{Expanded Measurement}$), but does not imply that latent utility is directly observed ($\text{AI} \not\Rightarrow \text{Latent Utility Observed}$). The AI signal expands the researcher's ability to measure preferences because it adds incremental information, but it does not convert latent utility into a directly observed object.

Proposition 2. *AI-Inferred Utility Is Not Invariant to Ordinal Transformations.*

Proof. Suppose utility is ordinal. Then any strictly increasing transformation $g(\cdot)$ represents the same preference ordering:

$$u_{ijt}^* \succ u_{ikt}^* \iff g(u_{ijt}^*) \succ g(u_{ikt}^*) \quad (20)$$

for any strictly increasing function g .

However, an AI-generated score $\hat{u}_{ijt}^{\text{AI}}$ is cardinal only relative to its own scale. Therefore, unless the scale is externally normalized or calibrated, differences such as:

$$\hat{u}_{ijt}^{\text{AI}} - \hat{u}_{ikt}^{\text{AI}} \quad (21)$$

cannot be interpreted as latent cardinal utility differences.

Ordinal preferences require only order preservation. If $u_{ijt}^* > u_{ikt}^*$ and g is strictly increasing, then $g(u_{ijt}^*) > g(u_{ikt}^*)$. Thus both u^* and $g(u^*)$ represent the same ranking. But cardinal differences are not invariant under monotonic transformations. In general:

$$u_{ijt}^* - u_{ikt}^* \neq g(u_{ijt}^*) - g(u_{ikt}^*) \quad (22)$$

For example, if $g(u) = \log(u)$, then:

$$g(u_a) - g(u_b) = \log(u_a) - \log(u_b) = \log\left(\frac{u_a}{u_b}\right) \neq u_a - u_b \quad (23)$$

Therefore, an AI-generated utility score cannot be treated as a latent cardinal utility scale unless additional normalization or calibration assumptions are imposed. $\square$

Interpretation: The researcher must not claim that the AI has recovered cardinal utility. Instead, the AI produces a cardinal-looking proxy that requires normalization and validation before any cardinal interpretation can be justified. It is a proxy, not an absolute psychological measure of preference.

4.3 Discrete Choice and the Random Utility Model

We extend the framework to a discrete choice environment by modeling consumer choices using a Random Utility Model (RUM) (McFadden 1974). Suppose individual i faces a set of alternatives

$j \in \{1, \dots, J\}$. The latent utility U_{ij} derived by individual i from alternative j is decomposed into a systematic component V_{ij} and an idiosyncratic random shock ε_{ij} :

$$U_{ij} = V_{ij} + \varepsilon_{ij} \quad (24)$$

The systematic component V_{ij} is a function of alternative characteristics X_{ij} , price p_j , and the calibrated AI-inferred utility proxy $\tilde{u}_{ij}^{\text{AI}}$:

$$V_{ij} = X_{ij}'\beta - \lambda p_j + \rho \tilde{u}_{ij}^{\text{AI}} \quad (25)$$

where β is a vector of preference weights, $\lambda > 0$ is price sensitivity, and ρ is the coefficient on the AI-inferred utility.

Assuming that the random shocks ε_{ij} are independently and identically distributed according to a Type I Extreme Value distribution, the probability P_{ij} that individual i chooses alternative j is given by the standard multinomial logit formula:

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{k=1}^J \exp(V_{ik})} \quad (26)$$

Within this model, the parameter ρ measures the marginal effect of the AI-inferred utility on the probability of choosing an alternative. If $\rho > 0$, this indicates that the AI-inferred utility contains incremental predictive content regarding consumer choice behavior beyond what is captured by standard observables X_{ij} and price p_j . Crucially, $\rho > 0$ does not imply that the AI has directly observed or recovered the consumer's latent utility. Rather, it indicates that the AI proxy is a statistically informative signal that correlates with the systematic determinants of choice.

4.4 Scale and Location Normalization in Discrete Choice

A fundamental property of discrete choice models is that choices are determined solely by ordinal utility differences, which makes the parameters of the model identified only up to location and scale. We formalize this in the following proposition.

Proposition 3. *Scale and Location Normalization in Discrete Choice. AI-inferred utility in discrete choice models is identified only up to location and scale; therefore, interpersonal cardinal comparisons require external normalization or experimental calibration.*

Proof. Let the latent utility of alternative j for individual i be $U_{ij} = V_{ij} + \varepsilon_{ij}$, where $V_{ij} = X'_{ij}\beta + \rho\tilde{u}^{\text{AI}}_{ij}$. Individual i chooses alternative j if and only if $U_{ij} \geq U_{ik}$ for all $k \neq j$.

Consider a positive affine transformation of the utility function:

$$U'_{ij} = \sigma U_{ij} + \alpha_i = \sigma(V_{ij} + \varepsilon_{ij}) + \alpha_i \quad (27)$$

where $\sigma > 0$ is a scale factor and $\alpha_i \in \mathbb{R}$ is an individual-specific location shift. Under this transformation, the choice probability P'_{ij} is:

$$P'_{ij} = \mathbb{P}(U'_{ij} \geq U'_{ik} \forall k) = \mathbb{P}(\sigma U_{ij} + \alpha_i \geq \sigma U_{ik} + \alpha_i \forall k) \quad (28)$$

Subtracting α_i and dividing by $\sigma > 0$ yields:

$$P'_{ij} = \mathbb{P}(U_{ij} \geq U_{ik} \forall k) = P_{ij} \quad (29)$$

Thus, choice behavior is invariant to any positive affine transformation of the underlying utility scale.

As a consequence, the absolute magnitude of the parameter ρ (and indeed the scale of $\tilde{u}^{\text{AI}}_{ij}$) cannot be identified from choice data alone; it is always relative to the scale of the error term ε_{ij} (which is normalized to $\sigma_\varepsilon^2 = \pi^2/6$ in the multinomial logit model). Therefore, interpersonal comparisons of cardinal utility using AI-inferred utility across different individuals or cohorts are invalid unless an external scale normalization or experimental calibration (e.g., matching the AI scores to a willingness-to-pay metric or expected utility lottery choices) is explicitly imposed. $\square$

4.5 Measurement Error and Econometric Bias

Proposition 4. *Classical Measurement Error Produces Attenuation Bias.*

Proof. Suppose the true model is:

$$Y_i = \alpha + \beta u_i^* + \varepsilon_i \quad (30)$$

but u_i^* is not observed. The researcher instead observes:

$$\hat{u}_i^{\text{AI}} = u_i^* + \eta_i \quad (31)$$

Assume:

$$\mathbb{E}[\eta_i] = 0, \quad \text{Cov}(u_i^*, \eta_i) = 0, \quad \text{Cov}(\varepsilon_i, u_i^*) = 0, \quad \text{Cov}(\varepsilon_i, \eta_i) = 0 \quad (32)$$

Then the probability limit of the OLS coefficient from regressing Y_i on $\hat{u}_i^{\text{AI}}$ is:

$$\text{plim}(\hat{\beta}_{\text{OLS}}) = \beta \frac{\text{Var}(u_i^*)}{\text{Var}(u_i^*) + \text{Var}(\eta_i)} \quad (33)$$

Thus, if $\text{Var}(\eta_i) > 0$, then:

$$|\text{plim}(\hat{\beta}_{\text{OLS}})| < |\beta| \quad (34)$$

The OLS coefficient is:

$$\hat{\beta}_{\text{OLS}} = \frac{\text{Cov}(Y_i, \hat{u}_i^{\text{AI}})}{\text{Var}(\hat{u}_i^{\text{AI}})} \quad (35)$$

Using the true model $Y_i = \alpha + \beta u_i^* + \varepsilon_i$ and the observed proxy $\hat{u}_i^{\text{AI}} = u_i^* + \eta_i$, we get:

$$\text{Cov}(Y_i, \hat{u}_i^{\text{AI}}) = \text{Cov}(\beta u_i^* + \varepsilon_i, u_i^* + \eta_i) \quad (36)$$

Expanding this covariance:

$$\text{Cov}(Y_i, \hat{u}_i^{\text{AI}}) = \beta \text{Cov}(u_i^*, u_i^*) + \beta \text{Cov}(u_i^*, \eta_i) + \text{Cov}(\varepsilon_i, u_i^*) + \text{Cov}(\varepsilon_i, \eta_i) \quad (37)$$

By the assumptions $\text{Cov}(u_i^*, \eta_i) = 0$, $\text{Cov}(\varepsilon_i, u_i^*) = 0$, and $\text{Cov}(\varepsilon_i, \eta_i) = 0$, we have:

$$\text{Cov}(Y_i, \hat{u}_i^{\text{AI}}) = \beta \text{Var}(u_i^*) \quad (38)$$

The variance of the observed proxy is:

$$\text{Var}(\hat{u}_i^{\text{AI}}) = \text{Var}(u_i^* + \eta_i) = \text{Var}(u_i^*) + \text{Var}(\eta_i) + 2\text{Cov}(u_i^*, \eta_i) \quad (39)$$

Since $\text{Cov}(u_i^*, \eta_i) = 0$, this simplifies to:

$$\text{Var}(\hat{u}_i^{\text{AI}}) = \text{Var}(u_i^*) + \text{Var}(\eta_i) \quad (40)$$

Substituting these back yields:

$$\text{plim}(\hat{\beta}_{\text{OLS}}) = \beta \frac{\text{Var}(u_i^*)}{\text{Var}(u_i^*) + \text{Var}(\eta_i)} \quad (41)$$

which is attenuated toward zero whenever $\text{Var}(\eta_i) > 0$. $\square$

Interpretation: When AI-inferred utility is used as a direct substitute for latent utility, classical measurement error generates attenuation bias. Hence, AI-based utility scores require calibration before structural interpretation. Direct plug-in without calibration will systematically understate the latent economic effect of utility on behavior.

Proposition 5. *Non-Classical AI Measurement Error Can Bias Estimates in Any Direction.*

Proof. Suppose $\hat{u}_i^{\text{AI}} = u_i^* + \eta_i$, but unlike the classical case, assume:

$$\text{Cov}(u_i^*, \eta_i) \neq 0 \quad \text{or} \quad \text{Cov}(\varepsilon_i, \eta_i) \neq 0 \quad (42)$$

Then the OLS coefficient from regressing Y_i on $\hat{u}_i^{\text{AI}}$ is not necessarily attenuated toward zero and may be biased upward, downward, or even change sign.

Again:

$$\hat{\beta}_{\text{OLS}} = \frac{\text{Cov}(Y_i, \hat{u}_i^{\text{AI}})}{\text{Var}(\hat{u}_i^{\text{AI}})} \quad (43)$$

Using $Y_i = \alpha + \beta u_i^* + \varepsilon_i$ and $\hat{u}_i^{\text{AI}} = u_i^* + \eta_i$, we obtain:

$$\text{Cov}(Y_i, \hat{u}_i^{\text{AI}}) = \text{Cov}(\beta u_i^* + \varepsilon_i, u_i^* + \eta_i) = \beta \text{Var}(u_i^*) + \beta \text{Cov}(u_i^*, \eta_i) + \text{Cov}(\varepsilon_i, u_i^*) + \text{Cov}(\varepsilon_i, \eta_i) \quad (44)$$

Assuming $\text{Cov}(\varepsilon_i, u_i^*) = 0$, this becomes:

$$\text{Cov}(Y_i, \hat{u}_i^{\text{AI}}) = \beta \text{Var}(u_i^*) + \beta \text{Cov}(u_i^*, \eta_i) + \text{Cov}(\varepsilon_i, \eta_i) \quad (45)$$

The denominator is:

$$\text{Var}(\hat{u}_i^{\text{AI}}) = \text{Var}(u_i^*) + \text{Var}(\eta_i) + 2\text{Cov}(u_i^*, \eta_i) \quad (46)$$

Therefore, the probability limit of the OLS estimator is:

$$\text{plim}(\hat{\beta}_{\text{OLS}}) = \frac{\beta \text{Var}(u_i^*) + \beta \text{Cov}(u_i^*, \eta_i) + \text{Cov}(\varepsilon_i, \eta_i)}{\text{Var}(u_i^*) + \text{Var}(\eta_i) + 2\text{Cov}(u_i^*, \eta_i)} \quad (47)$$

Because the numerator contains terms whose signs are unrestricted, the bias can be positive, negative, or sign-reversing. $\square$

Interpretation: This is a stronger and more realistic concern than classical attenuation. Be-

cause AI errors are often non-classical—correlating with language patterns, prices, demographics, or product categories—the bias induced by direct use of AI-inferred utility is not guaranteed to be attenuation bias. This motivates calibration and robustness analysis.

4.6 Calibration and Bounded Welfare Regret

To solve these measurement error problems, we apply the debiasing framework of Ludwig, Mullainathan, and Rambachan (2025). The researcher collects a human validation sample V containing human validation benchmark scores u_i^H , where $u_i^H = u_i^* + \nu_i$, and ν_i is human-labeling error.

Proposition 6. *Calibration Weakly Improves Measurement Accuracy.*

Proof. Let the calibrated AI utility be defined as the linear projection:

$$\tilde{u}_i^{\text{AI}} = \hat{a} + \hat{b}\hat{u}_i^{\text{AI}} \quad (48)$$

where $(\hat{a}, \hat{b})$ solve:

$$(\hat{a}, \hat{b}) = \arg \min_{a,b} \mathbb{E}[(u_i^H - a - b\hat{u}_i^{\text{AI}})^2] \quad (49)$$

Then:

$$\mathbb{E}[(u_i^H - \tilde{u}_i^{\text{AI}})^2] \leq \mathbb{E}[(u_i^H - \hat{u}_i^{\text{AI}})^2] \quad (50)$$

The calibrated predictor $\tilde{u}_i^{\text{AI}} = \hat{a} + \hat{b}\hat{u}_i^{\text{AI}}$ is chosen to minimize the mean square error $\mathbb{E}[(u_i^H - a - b\hat{u}_i^{\text{AI}})^2]$ over all linear transformations of $\hat{u}_i^{\text{AI}}$. The uncalibrated AI measure $\hat{u}_i^{\text{AI}}$ is a special case of this class with $a = 0$ and $b = 1$. Therefore, the minimum over the full class of linear transformations cannot be larger than the loss achieved by this special case.

Thus:

$$\mathbb{E}[(u_i^H - \tilde{u}_i^{\text{AI}})^2] = \min_{a,b} \mathbb{E}[(u_i^H - a - b\hat{u}_i^{\text{AI}})^2] \leq \mathbb{E}[(u_i^H - \hat{u}_i^{\text{AI}})^2] \quad (51)$$

□

Interpretation: Calibration is not an optional cosmetic step. It weakly improves the accuracy of AI-inferred utility relative to a validation benchmark and provides a scale correction necessary for structural interpretation.

The debiased estimator of the causal parameter is constructed as:

$$\hat{\beta}^{\text{debiased}} = \hat{\beta} - \hat{\lambda}_{\Delta|W} \quad (52)$$

where $\hat{\lambda}_{\Delta|W}$ is estimated by regressing the proxy discrepancy $\Delta_i = \hat{u}_i^{\text{AI}} - u_i^H$ on the covariate W_i within the validation sample.

Under the assumption of **no training leakage**, Ludwig et al. (2025) show that $\hat{\beta}^{\text{debiased}}$ is consistent and asymptotically normal. The asymptotic variance of the debiased estimator is given by:

$$\text{Var}(\hat{\beta}^{\text{debiased}}) = \frac{1}{N_R} \sigma_W^{-4} \left(\frac{1 - \rho_p}{\rho_p} \sigma_X^2 + 2\sigma_{XZ} + \frac{1 - \rho_v}{\rho_v} \sigma_Z^2 \right) \quad (53)$$

where N_R is the total number of observations, ρ_p is the proportion of observations in the estimation sample, ρ_v is the sampling ratio, $X_i = W_i \hat{u}_i^{\text{AI}}$, $Z_i = W_i \Delta_i$, and σ_W^2 is the variance of the covariate W_i .

Proposition 7. *Calibration Does Not Prove Recovery of Latent Utility.*

Proof. Even if $\mathbb{E}[(u_i^H - \tilde{u}_i^{\text{AI}})^2] \leq \mathbb{E}[(u_i^H - \hat{u}_i^{\text{AI}})^2]$, it does not follow that $\tilde{u}_i^{\text{AI}} = u_i^*$ unless $\nu_i = 0$ and the validation measure u_i^H itself perfectly equals the latent utility.

By definition, $u_i^H = u_i^* + \nu_i$. Calibration minimizes the distance to u_i^H , not directly to u_i^* . If $\nu_i \neq 0$, then even perfect prediction of u_i^H implies:

$$\tilde{u}_i^{\text{AI}} = u_i^H = u_i^* + \nu_i \quad (54)$$

which differs from u_i^* whenever $\nu_i \neq 0$. Therefore, calibration improves alignment with the human validation benchmark, but does not logically prove recovery of the underlying latent utility. $\square$

Interpretation: Calibration improves construct alignment relative to a validation benchmark, but it does not eliminate the philosophical and econometric gap between measured utility and latent utility.

Proposition 8. *Bounded AI Measurement Error Implies Bounded Allocation Loss.*

Proof. Suppose the AI-calibrated utility satisfies:

$$|\tilde{u}_{ij}^{\text{AI}} - u_{ij}^*| \leq \epsilon \quad (55)$$

for all j . Let the optimal choice be:

$$j^* = \arg \max_j u_{ij}^* \quad (56)$$

and the AI-based choice be:

$$\hat{j} = \arg \max_j \tilde{u}_{ij}^{\text{AI}} \quad (57)$$

Then the utility loss from choosing $\hat{j}$ instead of j^* is bounded by:

$$u_{ij^*}^* - u_{i\hat{j}}^* \leq 2\epsilon \quad (58)$$

Since $\hat{j}$ maximizes AI-calibrated utility:

$$\tilde{u}_{i\hat{j}}^{\text{AI}} \geq \tilde{u}_{ij^*}^{\text{AI}} \quad (59)$$

By the bounded error assumption:

$$u_{i\hat{j}}^* \geq \tilde{u}_{i\hat{j}}^{\text{AI}} - \epsilon \quad (60)$$

Also:

$$\tilde{u}_{ij^*}^{\text{AI}} \geq u_{ij^*}^* - \epsilon \quad (61)$$

Combining these inequalities:

$$u_{i\hat{j}}^* \geq \tilde{u}_{i\hat{j}}^{\text{AI}} - \epsilon \geq \tilde{u}_{ij^*}^{\text{AI}} - \epsilon \geq (u_{ij^*}^* - \epsilon) - \epsilon = u_{ij^*}^* - 2\epsilon \quad (62)$$

Rearranging terms yields:

$$u_{ij^*}^* - u_{i\hat{j}}^* \leq 2\epsilon \quad (63)$$

□

Interpretation: AI-based allocation need not require perfect utility recovery. If calibrated AI utility approximates latent utility within a bounded error, the welfare loss from AI-based allocation is itself bounded. This is a powerful result for market designers: operational efficacy does not require resolving metaphysical measurement debates.

4.7 Causal Identification Boundaries and Empirical Designs

To establish the utility of AI proxies in causal settings, we contrast prediction models with traditional causal identification designs.

Proposition 9. *Better Measurement Does Not Solve Causal Identification.*

Proof. Suppose the causal model is:

$$Y_i = \tau D_i + \theta u_i^* + X_i' \gamma + \varepsilon_i \quad (64)$$

where D_i is treatment, Y_i is outcome, and τ is the causal effect of interest. Even if AI perfectly recovers utility (such that $\tilde{u}_i^{\text{AI}} = u_i^*$), OLS estimation of τ is biased if:

$$\text{Cov}(D_i, \varepsilon_i \mid X_i, u_i^*) \neq 0 \quad (65)$$

Suppose the researcher estimates:

$$Y_i = \tau D_i + \theta \tilde{u}_i^{\text{AI}} + X_i' \gamma + \varepsilon_i \quad (66)$$

If $\tilde{u}_i^{\text{AI}} = u_i^*$, then utility is perfectly controlled for. However, the OLS estimator of τ is consistent only if:

$$\mathbb{E}[\varepsilon_i \mid D_i, X_i, u_i^*] = 0 \quad \text{or} \quad \text{Cov}(D_i, \varepsilon_i \mid X_i, u_i^*) = 0 \quad (67)$$

If instead $\text{Cov}(D_i, \varepsilon_i \mid X_i, u_i^*) \neq 0$, then treatment assignment remains endogenous. Therefore:

$$\text{plim}(\hat{\tau}_{\text{OLS}}) \neq \tau \quad (68)$$

even under perfect measurement of utility. □

Interpretation: Measurement improvement is not identification. Agentic AI may reduce measurement error, but it does not eliminate omitted-variable bias, simultaneity, selection, or treatment endogeneity.

Proposition 10. *Randomization Identifies the Treatment Effect Independently of AI.*

Proof. Let potential outcomes be $Y_i(1)$ and $Y_i(0)$, and treatment $D_i \in \{0, 1\}$. If treatment is

randomly assigned:

$$D_i \perp (Y_i(1), Y_i(0), u_i^*, \tilde{u}_i^{\text{AI}}) \quad (69)$$

then the average treatment effect τ is identified by:

$$\tau = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \quad (70)$$

Observed outcome is:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \quad (71)$$

By random assignment:

$$\mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i(1)] \quad \text{and} \quad \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i(0)] \quad (72)$$

Therefore:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau \quad (73)$$

□

Interpretation: The source of causal identification is randomization, not AI. AI can improve the measurement of heterogeneous treatment effects $\tau(\tilde{u})$, but it is the experimental design that establishes causality.

Proposition 11. *Instrumental Variables Identify Causal Effects Only Under Validity Conditions.*

Proof. Suppose treatment D_i is endogenous in:

$$Y_i = \tau D_i + \theta \tilde{u}_i^{\text{AI}} + X_i' \gamma + \varepsilon_i \quad (74)$$

Let Z_i be an instrument satisfying:

1. Relevance: $\text{Cov}(Z_i, D_i \mid X_i, \tilde{u}_i^{\text{AI}}) \neq 0$
2. Exogeneity: $\text{Cov}(Z_i, \varepsilon_i \mid X_i, \tilde{u}_i^{\text{AI}}) = 0$
3. Exclusion Restriction: Z_i affects Y_i only through D_i .

Then τ is identified by instrumental variables (Imbens & Angrist 1994).

The moment condition implied by exogeneity is:

$$\mathbb{E}[Z_i(Y_i - \tau D_i - \theta \tilde{u}_i^{\text{AI}} - X_i' \gamma)] = 0 \quad (75)$$

Substituting the structural equation $Y_i - \tau D_i - \theta \tilde{u}_i^{\text{AI}} - X_i' \gamma = \varepsilon_i$, we obtain:

$$\mathbb{E}[Z_i \varepsilon_i] = 0 \quad (76)$$

By relevance, Z_i generates variation in D_i . Therefore, the variation in D_i induced by Z_i is exogenous with respect to ε_i , allowing identification of τ . In the just-identified scalar case:

$$\tau_{\text{IV}} = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, D_i)} \quad (77)$$

after residualizing with respect to X_i and $\tilde{u}_i^{\text{AI}}$. $\square$

Interpretation: Agentic AI may assist in discovering candidate instruments (e.g., searching news databases), but the validity of an instrument remains a theoretical and institutional claim, not an output of the AI model itself.

Proposition 12. *Difference-in-Differences Requires Parallel Trends, Not Better Text Measurement.*

Proof. Let groups be $g \in \{0, 1\}$ and time be $t \in \{0, 1\}$. The DiD estimator is:

$$\hat{\tau}_{\text{DiD}} = (\mathbb{E}[Y_{1,1}] - \mathbb{E}[Y_{1,0}]) - (\mathbb{E}[Y_{0,1}] - \mathbb{E}[Y_{0,0}]) \quad (78)$$

It identifies the treatment effect if the untreated potential outcomes satisfy parallel trends:

$$\mathbb{E}[Y_{1,1}(0) - Y_{1,0}(0)] = \mathbb{E}[Y_{0,1}(0) - Y_{0,0}(0)] \quad (79)$$

Observed treated-group change is:

$$\mathbb{E}[Y_{1,1}] - \mathbb{E}[Y_{1,0}] = \mathbb{E}[Y_{1,1}(1)] - \mathbb{E}[Y_{1,0}(0)] \quad (80)$$

Adding and subtracting $\mathbb{E}[Y_{1,1}(0)]$ yields:

$$\mathbb{E}[Y_{1,1}(1) - Y_{1,1}(0)] + \mathbb{E}[Y_{1,1}(0) - Y_{1,0}(0)] \quad (81)$$

The control-group change is:

$$\mathbb{E}[Y_{0,1}] - \mathbb{E}[Y_{0,0}] = \mathbb{E}[Y_{0,1}(0) - Y_{0,0}(0)] \quad (82)$$

Under parallel trends, subtracting the control-group change removes the trend term, leaving:

$$\hat{\tau}_{\text{DiD}} = \mathbb{E}[Y_{1,1}(1) - Y_{1,1}(0)] \quad (83)$$

□

Interpretation: AI-generated measures can enrich the outcome or covariate space in a DiD design, but they do not substitute for the parallel-trends assumption.

We complement these mathematical boundaries with traditional empirical identification designs in practice. For instance, in an RCT, treatment D_i is randomized, making it independent of the measurement error η_i ($\text{Cov}(D_i, \eta_i) = 0$), which mitigates selection bias but still requires calibration to address attenuation. In IV designs, AI can crawl text corpuses to suggest instruments, but qualitative proof of the exclusion restriction is mandatory. In RD designs, algorithms that estimate running variables (e.g., credit risk scores or sentiment indexes) must not exhibit rounding discontinuities that violate running variable density continuity. Finally, in DiD designs, stable prompt contexts are required to prevent artificial temporal shifts from violating the parallel trends assumption.

4.8 Main Theorem: Bounded Computability of Economic Preferences

Theorem 1 (Bounded Computability of Economic Preferences).

Statement of the Theorem. Under the following assumptions:

1. AI-inferred utility contains incremental information:

$$\mathbb{E}[u_i^* \mid X_i, p_i, \hat{u}_i^{\text{AI}}] \neq \mathbb{E}[u_i^* \mid X_i, p_i] \quad (84)$$

2. AI measurement error is finite:

$$\text{Var}(\eta_i) < \infty \quad (85)$$

3. A validation sample exists:

$$u_i^H = u_i^* + \nu_i \quad (86)$$

4. Calibrated utility satisfies bounded error:

$$|\tilde{u}_i^{\text{AI}} - u_i^*| \leq \epsilon \quad (87)$$

5. Causal identification is supplied externally through RCT, IV, DiD, or RD.

Then agentic AI expands the computability of economic preferences by improving measurement and allocation, but it does not directly reveal latent utility and does not independently identify causal effects.

Formally:

$$\mathbb{E}[(u_i^* - \mathbb{E}[u_i^* | X_i, p_i, \hat{u}_i^{\text{AI}}])^2] \leq \mathbb{E}[(u_i^* - \mathbb{E}[u_i^* | X_i, p_i])^2] \quad (88)$$

and:

$$u_{i,\text{best}}^* - u_{i,\text{AI-choice}}^* \leq 2\epsilon \quad (89)$$

but:

$$\text{Cov}(D_i, \varepsilon_i | X_i, \tilde{u}_i^{\text{AI}}) \neq 0 \Rightarrow \text{plim}(\hat{\tau}_{\text{OLS}}) \neq \tau \quad (90)$$

□

Proof of the Theorem. The first inequality follows from Proposition 1: adding AI-inferred utility to the information set weakly improves mean-square prediction of latent utility.

The allocation bound follows from Proposition 8: if calibrated AI utility approximates latent utility within bounded error ϵ , then the regret from AI-based allocation is bounded by 2ϵ .

The causal limitation follows from Proposition 9: even perfect measurement of utility does not remove treatment endogeneity when treatment remains correlated with unobserved determinants of the outcome.

Therefore, agentic AI expands measurement and operational computability, but does not dissolve the theoretical distinction between measured proxy and latent utility, nor the econometric distinction between prediction and causal identification. □

This theorem should be interpreted as a bounded measurement result, not as a metaphysical claim about utility. It shows that AI can improve the operational approximation of preference-

related variables when its signal contains incremental information, its error is bounded, and its scale is externally validated. It does not imply that utility becomes directly observable, interpersonally comparable, or causally identified. To illustrate the implications of the bounded-measurement theorem, Appendix B provides Monte Carlo simulations of measurement error, calibration, coefficient bias, random utility prediction, and bounded allocation regret under classical and non-classical AI proxy error structures. The simulations are illustrative rather than empirical: they do not assume that benchmark accuracy rates from existing LLM studies are direct estimates of utility-scale error.

Interpretation: The central contribution of this paper is not the claim that agentic AI makes latent utility directly observable. Rather, the paper establishes a bounded-computability result. If AI-inferred utility contains information about latent preferences beyond standard observables, then it expands the measurement frontier. If its measurement error can be calibrated using a validation sample, then it can improve prediction and reduce allocation loss. However, AI-inferred utility remains a proxy, not the latent object itself. Moreover, measurement improvement does not imply causal identification. Causal claims still require external sources of variation such as randomization, instrumental variables, difference-in-differences, or regression discontinuity designs.

5 Falsification of Bounded Computability

The bounded-computability framework presented in Theorem 1 relies on a set of precise identifying assumptions. Consequently, the claim of bounded computability is falsifiable. The core hypothesis of bounded computability fails under any of the following empirical or econometric conditions:

1. **Zero Incremental Information:** The bounded-computability claim fails if AI-inferred utility adds no information beyond standard observables. Formally, this occurs if:

$$\mathbb{E}[u_i^* \mid X_i, p_i, \hat{u}_i^{\text{AI}}] = \mathbb{E}[u_i^* \mid X_i, p_i] \quad (91)$$

In this case, the AI proxy has no predictive or measurement value, and the information expansion result collapses.

2. **Unbounded or Systematically Biased Measurement Error:** If the measurement

error variance is infinite ($\text{Var}(\eta) = \infty$), or if the error η is systematically biased in a way that correlates with unobserved variables and cannot be bounded, then the welfare regret from AI-based allocation cannot be bounded. That is, if ϵ is unbounded, the allocation regret 2ϵ becomes arbitrarily large.

3. **Non-Representative Validation Sample:** The debiasing framework relies on the assumption that the human validation sample V is representative of the estimation sample R . If V is selected non-randomly or suffers from selection bias (e.g., human-labeled benchmarks represent a non-representative subgroup), then the estimated measurement error bias $\hat{\lambda}_{\Delta|W}$ will be inconsistent for the true bias, invalidating the causal debiasing.
4. **Training Leakage:** If the machine learning model’s training set contains the evaluation, estimation, or validation data (data leakage), the AI-inferred utility $\hat{u}_i^{\text{AI}}$ will exhibit artificially low error within the sample but fail to generalize. Causal inference under the Ludwig et al. (2025) framework requires independence of the error terms across samples, which is violated by training leakage.
5. **Absence of Causal Identification:** Even if the AI proxy is highly accurate, any causal claim regarding policy changes or treatment effects fails if a valid identification strategy (such as an RCT, IV, RD, or DiD) is absent. A pure prediction model cannot establish causality.

6 Applications and Extensions

The following applications illustrate how the bounded-measurement framework can be used in adjacent economic domains. They are not part of the core identification theorem. The core contribution of the paper remains the distinction between AI-based measurement improvement, latent utility recovery, and causal identification. The applications below show how this distinction matters for productivity, knowledge dynamics, pricing, and production theory.

6.1 Micro-Productivity and Macroeconomic Growth

The macroeconomic impact of agentic AI is characterized by a tension between rapid productivity gains observed at the micro task level and modest projections for aggregate economic growth.

6.1.1 Micro-level Empirical Evidence

Recent empirical studies show that generative AI significantly boosts individual worker productivity, particularly for low-skilled or newer employees:

- **Customer Support (Brynjolfsson, Li & Raymond 2025):** In a study of 5,172 customer support agents, the introduction of a conversational AI assistant increased productivity (resolved issues per hour) by about 15% on average. Crucially, the productivity of low-skilled and novice workers increased by 34%, while the impact on highly skilled workers was negligible. The AI assistant acted as a mechanism for capturing and distributing tacit knowledge from top performers to the rest of the workforce.
- **Writing Tasks (Noy & Zhang 2023):** In an experimental setup with 453 college-educated professionals performing mid-level writing tasks, ChatGPT reduced task completion time by 40% while increasing output quality (as graded by external evaluators) by 18%.
- **Software Development (Peng et al. 2023):** In a randomized trial with 70 professional programmers, developers using GitHub Copilot completed a standardized coding task 55.8% faster than the control group.

6.1.2 Macroeconomic TFP Growth: The Acemoglu (2024) Framework

Despite these microeconomic gains, Daron Acemoglu (2024) argues that the aggregate impact of AI on Total Factor Productivity (TFP) will be modest over the next decade.

Acemoglu uses a task-based model of production. Let output be produced by combining a continuum of tasks $i \in [0, 1]$:

$$\ln Y = \int_0^1 \ln q(i) di \quad (92)$$

where $q(i)$ is the service level of task i . AI can automate a share s_{tasks} of these tasks. According to Hulten's Theorem, the first-order effect of a technological change on TFP is equal to the share of tasks automated multiplied by the average cost savings at the task level ($\bar{\pi}$):

$$d \ln TFP \approx \bar{\pi} \times s_{tasks} \quad (93)$$

Acemoglu estimates that:

1. About 19.9% of tasks are exposed to AI, but only 23% of these exposed tasks can be automated profitably, yielding a share of automated tasks of $s_{\text{tasks}} \approx 0.199 \times 0.23 \approx 4.6\%$.
2. For these automated tasks, the average task-level labor cost savings is approximately 27%, which translates to an overall cost savings of $\bar{\pi} \approx 14.4\%$ (accounting for the labor share of costs in these sectors).
3. This yields a baseline aggregate TFP growth of:

$$d \ln \text{TFP} \approx s_{\text{tasks}} \times \bar{\pi} \approx 0.046 \times 0.144 \approx 0.66\% \text{ over 10 years} \quad (94)$$

which is roughly 0.066% annually. Under a more conservative scenario that accounts for the difficulty of automating “hard-to-learn” tasks, the TFP gain falls below 0.53% over 10 years.

This modest projection arises because many high-value, complex tasks (requiring social interaction, physical dexterity, or open-ended reasoning) cannot be easily automated by current AI systems.

6.1.3 Heterogeneous Agents: DSGE and HANK Models

To understand how these productivity gains are distributed, modern macroeconomics incorporates AI agents into Dynamic Stochastic General Equilibrium (DSGE) and Heterogeneous Agent New Keynesian (HANK) models.

In a standard representative-agent DSGE model (Smets & Wouters 2007), the household maximizes lifetime utility:

$$\max \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(C_t, N_t) \quad (95)$$

subject to the budget constraint:

$$C_t + \frac{B_{t+1}}{R_t} = w_t N_t + B_t + T_t \quad (96)$$

yielding the classic Euler equation:

$$U_C(C_t) = \beta R_t \mathbb{E}_t[U_C(C_{t+1})] \quad (97)$$

HANK models (Kaplan, Moll & Violante 2018) relax the representative-agent assumption by

introducing a continuum of households facing idiosyncratic income shocks and credit constraints. In this setting, the introduction of AI has two distribution channels:

- **Labor Income Channel:** AI increases the wages of low-skilled workers (as seen in Brynjolfsson et al. 2025), reducing income inequality and lowering the aggregate Marginal Propensity to Consume (MPC).
- **Capital Income Channel:** If AI technologies are owned by a small group of capital holders, the gains accrue as corporate profits. This increases wealth inequality and concentrates income among low-MPC households, potentially leading to aggregate demand deficiency.

6.2 The Dynamics of Knowledge Collapse

A major risk of the widespread adoption of agentic AI is **knowledge collapse** (Peterson 2024; Acemoglu, Lin, Ozdaglar & Siderius 2026). This section formalizes the dynamic feedback loop between human learning and AI recommendations.

6.2.1 The Dynamic Model

Consider a society where individuals face a sequence of decisions. Making an optimal decision requires information about the state of the world $\theta \sim \mathcal{N}(0, \sigma_0^2)$. Individuals can acquire information through two channels:

1. **Exploration (Human Effort):** Engaging in costly research and direct observation to obtain a private signal $s_i = \theta + v_i$, where $v_i \sim \mathcal{N}(0, \sigma_v^2)$. The cost of effort is $C(e_i) = \frac{1}{2}e_i^2$.
2. **Exploitation (AI Recommendation):** Querying an LLM agent that pools historical data to provide a recommendation $r_i = \theta + \eta_i$, where $\eta_i \sim \mathcal{N}(0, \sigma_{LLM}^2)$. Querying the AI is cheap, costing only $c \approx 0$.

The LLM agent’s accuracy depends on the historical accumulation of human exploration. The variance of the AI’s error at time t is an inverse function of the total human exploration effort in previous periods:

$$\sigma_{LLM,t}^2 = \frac{\kappa}{\sum_{\tau=1}^{t-1} E_\tau} \quad (98)$$

where $E_\tau = \sum_i e_{i,\tau}$ is the aggregate human effort at time τ .

When agentic AI becomes highly accurate (σ_{LLM}^2 is small), individuals substitute away from costly exploration toward cheap AI exploitation. In the limit, human effort $e_i \rightarrow 0$.

Proposition 13. *Knowledge Collapse Can Occur If AI Reduces Human Learning Incentives.*

Proof. Let human knowledge stock evolve as:

$$H_{t+1} = (1 - \delta_H)H_t + \phi e_t \quad (99)$$

where H_t is human knowledge, e_t is human learning effort, $\delta_H \in (0, 1)$ is the depreciation rate, and $\phi > 0$.

Suppose individuals choose effort to maximize their net payoff:

$$e_t = \arg \max_e [B(H_t, e, A_t) - C(e)] \quad (100)$$

Assume:

$$\frac{\partial^2 B}{\partial e \partial A} < 0 \quad (101)$$

meaning that AI adoption reduces the marginal benefit of human learning effort.

The first-order condition for optimal effort e_t^{opt} is:

$$\frac{\partial B}{\partial e}(H_t, e_t^{\text{opt}}, A_t) = C'(e_t^{\text{opt}}) \quad (102)$$

Differentiating implicitly with respect to AI intensity A_t yields:

$$\frac{\partial^2 B}{\partial e^2} \frac{\partial e_t^{\text{opt}}}{\partial A_t} + \frac{\partial^2 B}{\partial e \partial A} = C''(e_t^{\text{opt}}) \frac{\partial e_t^{\text{opt}}}{\partial A_t} \quad (103)$$

Rearranging terms:

$$\left(\frac{\partial^2 B}{\partial e^2} - C''(e_t^{\text{opt}}) \right) \frac{\partial e_t^{\text{opt}}}{\partial A_t} = - \frac{\partial^2 B}{\partial e \partial A} \quad (104)$$

Under second-order sufficiency conditions (concavity of the objective function), the term in parentheses is negative:

$$\frac{\partial^2 B}{\partial e^2} - C''(e_t^{\text{opt}}) < 0 \quad (105)$$

Since $\frac{\partial^2 B}{\partial e \partial A} < 0$, we have $-\frac{\partial^2 B}{\partial e \partial A} > 0$. Thus, solving for the derivative:

$$\frac{\partial e_t^{\text{opt}}}{\partial A_t} < 0 \quad (106)$$

So AI intensity reduces learning effort.

Human knowledge declines when $H_{t+1} < H_t$. Substituting the law of motion:

$$(1 - \delta_H)H_t + \phi e_t^{\text{opt}} < H_t \iff \phi e_t^{\text{opt}} < \delta_H H_t \iff e_t^{\text{opt}} < \frac{\delta_H}{\phi} H_t \quad (107)$$

which is satisfied if AI intensity A_t is sufficiently high such that optimal effort falls below the replacement threshold. $\square$

Interpretation: Knowledge collapse is a conditional dynamic outcome that occurs only when AI adoption reduces human learning effort below the replacement threshold required to maintain the knowledge stock.

6.2.2 Mathematical Formulation of PDF Shrinkage

As human exploration ceases, the AI is trained increasingly on its own generated outputs from previous periods. This recursive feedback loop alters the probability density function (PDF) of the shared knowledge base.

Let $p_0(x)$ be the true distribution of human knowledge. The LLM generative process can be modeled as a contraction mapping:

$$p_{t+1}(x) = \frac{p_t(x)^\gamma}{\int_{-\infty}^{\infty} p_t(y)^\gamma dy} \quad (108)$$

where $\gamma > 1$ represents the model's tendency to sample from the center of the distribution (temperature scaling or reinforcement learning from human feedback).

If the initial distribution of knowledge is normal, $X_0 \sim \mathcal{N}(\mu, \sigma_0^2)$, with PDF:

$$p_0(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma_0^2}\right) \quad (109)$$

then after t iterations, the PDF of the knowledge base remains normal:

$$p_t(x) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma_t^2}\right) \quad \text{where } \sigma_t^2 = \frac{\sigma_0^2}{\gamma^t} \quad (110)$$

As $t \rightarrow \infty$, the variance $\sigma_t^2 \rightarrow 0$, and the distribution collapses to a Dirac delta function centered at the mean μ .

Figure 3 illustrates this shrinkage of the knowledge distribution.

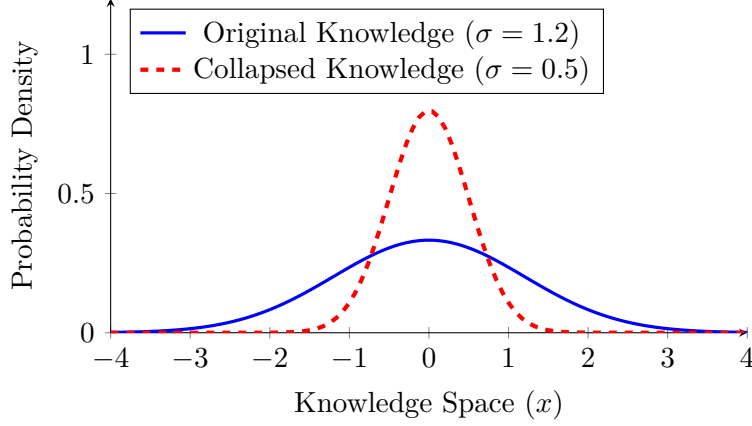


Figure 3: The Shrinkage of Knowledge Distribution under AI Feedback Loops

This concentration toward the mean eliminates the tails of the distribution. In economics, these tails represent rare, high-impact innovations, heterodox ideas, and entrepreneurial discoveries. The loss of these tails leads to intellectual stagnation and a reduction in the society’s capacity to adapt to new macroeconomic shocks.

6.2.3 Mitigation: Controlled Information Degradation

To prevent knowledge collapse, Acemoglu, Lin, Ozdaglar & Siderius (2026) suggest that platforms must intentionally introduce noise or limit the precision of AI recommendations—a policy known as **information garbling**.

Let the recommended signal be modified by adding a garbling term $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$:

$$\tilde{r}_i = r_i + \xi_i \tag{111}$$

By increasing the variance of the AI recommendation, the relative value of private exploration increases, incentivizing humans to continue generating the novel signals required to update the public database.

6.3 Algorithmic Pricing, Personalization, and Collusion

The implementation of agentic AI by firms has deep implications for market structure, consumer surplus, and antitrust policy.

6.3.1 Personalized Pricing and Consumer Surplus Capture

First-degree price discrimination occurs when a firm charges each customer their exact reservation price (willingness to pay). Historically, this was impossible due to search costs and asymmetric information. Today, LLM agents can analyze a consumer's browsing history, location, and textual inputs to estimate their marginal utility for a product in real time.

Consider a market with demand curve $P(Q) = A - BQ$ and marginal cost $MC = c$. Under perfect competition, price equals marginal cost ($P_c = c$), and Consumer Surplus (CS) is maximized. Under a single-price monopoly, the firm sets marginal revenue equal to marginal cost ($MR = MC$), charging a uniform price $P_m = (A + c)/2$, which creates a Deadweight Loss (DWL) representing lost economic efficiency.

If the monopolist uses an LLM agent to practice perfect personalized pricing, the firm charges consumer i exactly their reservation price $p_i = u_i^*(x)$. In this case:

- The firm expands output to the competitive level $Q^* = (A - c)/B$, eliminating deadweight loss.
- Total economic surplus is maximized, but **Consumer Surplus is reduced to zero**.
- The entire economic surplus is captured by the monopolist as producer surplus (PS).

Figure 4 illustrates this redistribution of welfare.

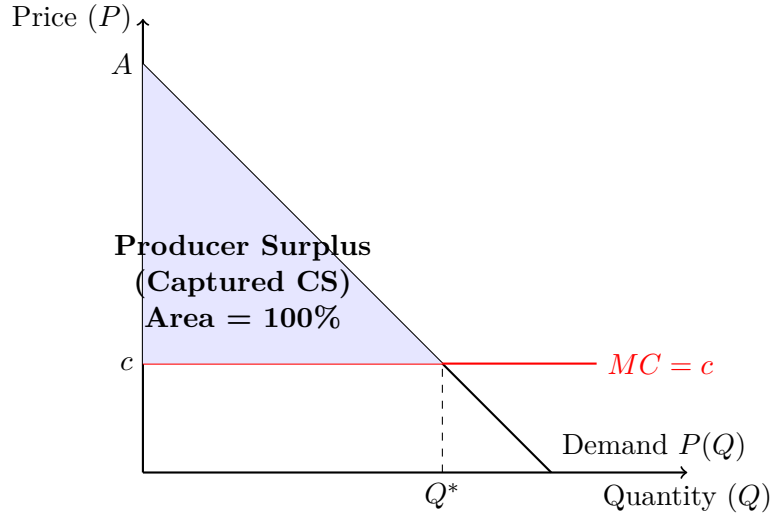


Figure 4: Welfare Under First-Degree Price Discrimination via AI Agents

Proposition 14. *AI-Augmented Utility Can Improve Pricing Decisions Under Demand Regularity.*

Proof. Suppose demand for product j is given by $q_j(p_j, \tilde{u}_j^{\text{AI}})$, where p_j is the price, $\tilde{u}_j^{\text{AI}}$ is the calibrated AI utility proxy, and demand satisfies:

$$\frac{\partial q_j}{\partial p_j} < 0 \quad \text{and} \quad \frac{\partial q_j}{\partial \tilde{u}_j^{\text{AI}}} > 0 \quad (112)$$

Firm profit is:

$$\pi_j(p_j) = (p_j - c_j)q_j(p_j, \tilde{u}_j^{\text{AI}}) \quad (113)$$

where c_j is the constant marginal cost.

Differentiating profit with respect to price p_j yields the first-order condition:

$$\frac{\partial \pi_j}{\partial p_j} = q_j(p_j, \tilde{u}_j^{\text{AI}}) + (p_j - c_j) \frac{\partial q_j}{\partial p_j} \quad (114)$$

Setting the first-order condition equal to zero at the optimal price:

$$q_j(p_j, \tilde{u}_j^{\text{AI}}) + (p_j - c_j) \frac{\partial q_j}{\partial p_j} = 0 \quad (115)$$

Dividing both sides by $q_j(p_j, \tilde{u}_j^{\text{AI}})$ yields:

$$1 + (p_j - c_j) \frac{1}{q_j} \frac{\partial q_j}{\partial p_j} = 0 \quad (116)$$

Multiplying and dividing the second term by p_j :

$$1 + \frac{p_j - c_j}{p_j} \left(\frac{\partial q_j}{\partial p_j} \frac{p_j}{q_j} \right) = 0 \quad (117)$$

Defining the price elasticity of demand as $\varepsilon_j = \frac{\partial q_j}{\partial p_j} \frac{p_j}{q_j}$, we obtain:

$$1 + \frac{p_j - c_j}{p_j} \varepsilon_j = 0 \quad (118)$$

Solving for the Lerner Index yields the optimal pricing rule:

$$\frac{p_j - c_j}{p_j} = -\frac{1}{\varepsilon_j} \quad (119)$$

□

Interpretation: In food and beverage (F&B) and other retail applications, AI-inferred utility

is operationally valuable because it can improve demand estimation and pricing decisions, not because it directly reveals psychological utility. The Lerner Index shows that pricing depends on the price elasticity of demand, which can be estimated more precisely by conditioning on the AI-augmented utility signal.

6.3.2 Algorithmic Tacit Collusion

A major concern for antitrust authorities is that pricing algorithms can learn to collude without explicit human communication.

In a repeated Bertrand-Nash game, firms choose prices simultaneously. If the game is repeated infinitely, firms can sustain monopoly prices using trigger strategies, provided the discount factor δ is sufficiently high:

$$\delta \geq \frac{P_m - P_c}{P_m - P_d} \quad (120)$$

where P_d is the payoff from deviating.

Calvano et al. (2020) demonstrated that Q-learning algorithms (a standard reinforcement learning method) consistently learn to collude when playing repeated pricing games. The algorithms start by testing different prices. Over time, they learn that cutting prices leads to retaliatory price wars by the competitor’s algorithm. Consequently, they converge to stable, supra-competitive prices.

Antitrust laws (such as Section 1 of the Sherman Act in the US or Article 101 of the TFEU in the EU) require proof of an agreement (collusion) to establish a violation. Because LLM pricing agents can coordinate prices tacitly through feedback loops without any explicit agreement (Calvano et al. 2020; Klein 2021; Fish, Gonczarowski & Shorrer 2024), they operate in a legal gray area that poses a severe challenge to regulators.

6.4 CES Production with Agentic Capital

In Section 6.1, we discussed the macroeconomic implications of AI on growth and productivity. In this subsection, we formalize the micro-foundations of AI as a production input. We model agentic AI as a distinct form of capital within a Constant Elasticity of Substitution (CES) production framework.

Proposition 15. *CES Production with Agentic Capital.*

Proof. Let aggregate output Y be produced using physical capital K , labor L , and agentic AI capital A according to the CES production function:

$$Y = \Omega [\alpha K^\rho + \beta L^\rho + \delta A^\rho]^{\frac{1}{\rho}} \quad (121)$$

where $\Omega > 0$ is a scale parameter (representing Total Factor Productivity), $\alpha, \beta, \delta > 0$ are the distribution parameters satisfying $\alpha + \beta + \delta = 1$, and $\rho \leq 1$ is the substitution parameter. The elasticity of substitution between inputs is:

$$\sigma = \frac{1}{1 - \rho} \quad (122)$$

We wish to derive the marginal product of agentic AI capital, $MP_A = \frac{\partial Y}{\partial A}$.

Let the term in brackets be defined as:

$$S = \alpha K^\rho + \beta L^\rho + \delta A^\rho \quad (123)$$

so that the production function can be written as $Y = \Omega S^{\frac{1}{\rho}}$. Differentiating Y with respect to A using the chain rule yields:

$$\frac{\partial Y}{\partial A} = \Omega \frac{1}{\rho} S^{\frac{1}{\rho}-1} \frac{\partial S}{\partial A} \quad (124)$$

Differentiating S with respect to A :

$$\frac{\partial S}{\partial A} = \delta \rho A^{\rho-1} \quad (125)$$

Substituting this derivative back into the expression for $\frac{\partial Y}{\partial A}$:

$$MP_A = \Omega \frac{1}{\rho} S^{\frac{1}{\rho}-1} (\delta \rho A^{\rho-1}) \quad (126)$$

Canceling the parameter ρ yields:

$$MP_A = \Omega \delta A^{\rho-1} S^{\frac{1}{\rho}-1} \quad (127)$$

Substituting back the full expression for S , we obtain the marginal product of agentic AI capital:

$$MP_A = \Omega \delta A^{\rho-1} [\alpha K^\rho + \beta L^\rho + \delta A^\rho]^{\frac{1}{\rho}-1} \quad (128)$$

□

Interpretation: Agentic AI can be modeled as a productive input, but its aggregate impact depends on its marginal product, adoption share, substitutability with labor, and task exposure. Modeling AI as a capital asset allows macroeconomists to study the dynamic paths of wages and return on capital during transition phases of automation.

Proposition 16. *Agentic AI Raises Output If Its Marginal Product Exceeds Its User Cost.*

Proof. Let the price of output be normalized to P . Let w be the wage rate of labor, r be the user cost of physical capital, and r_A be the rental price (user cost) of agentic AI capital. A representative firm chooses inputs K , L , and A to maximize profits:

$$\Pi = PY(K, L, A) - wL - rK - r_A A \quad (129)$$

Differentiating the profit function with respect to agentic AI capital A yields:

$$\frac{\partial \Pi}{\partial A} = P \frac{\partial Y}{\partial A} - r_A = P \cdot MP_A - r_A \quad (130)$$

A marginal increase in the adoption of agentic AI capital A increases the firm's profits if:

$$\frac{\partial \Pi}{\partial A} > 0 \iff P \cdot MP_A > r_A \quad (131)$$

At the interior optimum, the firm sets the marginal revenue product of agentic capital equal to its user cost:

$$P \cdot MP_A = r_A \quad (132)$$

□

Interpretation: AI adoption is profitable only when the value of its marginal product exceeds its cost. Firms will not adopt AI simply because it is technically feasible; the adoption decision is fundamentally governed by relative factor prices and task-level productivity gains.

7 Policy Recommendations and Conclusion

7.1 Econometric Standards and Robustness Protocol

To ensure the scientific validity of empirical research using AI-generated variables, we recommend that researchers, academic journals, and funding agencies adopt a rigorous robustness protocol consisting of the following standards:

1. **Mandatory Human Validation Sample:** Any study using AI-inferred variables (such as simulated economic preferences, sentiment scores, or text classifications) as primary independent or dependent variables must include a human-audited validation sample ($n_v \geq 250$) to estimate the measurement error profile and construct the calibrated proxy.
2. **Prompt Pre-Registration:** Prompt templates, system instructions, and temperature parameters must be pre-registered to prevent prompt-hacking or *p*-hacking through iterative post-hoc adjustments.
3. **Model-Version Pinning:** Researchers must pin the exact model version and API snapshot (e.g., `gpt-4-0314` rather than dynamic or updated model endpoints) to ensure reproducibility and prevent drift in inferred values.
4. **Leakage Audit:** Researchers must conduct a formal audit to verify whether the training or fine-tuning datasets of the utilized LLMs contain the evaluation, estimation, or validation samples (data leakage).
5. **Multiple Prompts (Robustness to Phrasing):** Studies must test the sensitivity of the AI-inferred variables to alternative phrasings, synonyms, and ordering of prompt instructions, reporting the variance of the resulting variables.
6. **Multiple Models:** Researchers must run the utility inference using multiple independent LLMs from different developers (e.g., OpenAI, Anthropic, Google) to ensure results are not model-specific artifacts.
7. **Placebo Prompts:** Researchers must include placebo or uninformative prompts (such as scrambled text, random reviews, or neutral fillers) to confirm that the AI proxy is extracting meaningful economic signals rather than fitting noise or responding to arbitrary syntactic structures.

8. **Bootstrap Standard Errors:** Studies must report bootstrapped standard errors that incorporate the additional estimation uncertainty introduced by the validation sample and calibration step.
9. **Sensitivity Analysis:** Researchers must conduct sensitivity analyses that trace how the causal parameter estimates change under variations in the calibration slope (b) and the error variance ($\text{Var}(\eta)$), particularly when the validation sample is small or subject to human-labeling noise.

7.2 Regulatory Antitrust Frameworks

To address the rise of algorithmic collusion and consumer surplus extraction, we propose:

1. **algorithmic Audit Trails:** Firms utilizing automated pricing agents must maintain immutable logs of their algorithm’s objective functions, reward matrices, and training histories.
2. **Consumer Data Portability:** To weaken first-degree price discrimination, consumers must have ownership of their behavioral data, allowing them to clear their history or transmit it to competing purchasing agents.
3. **Exploration Mandates:** For platforms utilizing recommendation algorithms, regulators should mandate a minimum share of randomized search results (information garbling) to maintain human discovery incentives and prevent knowledge collapse.

7.3 Conclusion

The integration of agentic AI into economic systems does not resolve the classical marginal utility computability dilemma. It does not eliminate the latent nature of utility, nor does it resolve the problem of interpersonal utility comparisons.

However, agentic AI represents a massive leap in **measurement and operational capability**. By acting as a sophisticated translator of unstructured human behavior, LLMs allow economists and market designers to approximate utility with unprecedented detail. The challenge for the economics profession is to combine this measurement power with the rigorous causal designs and human-validated methods that have defined the discipline since the Marginal Revolution.

Appendix B: Monte Carlo Illustration of Bounded Measurement and Calibration

B.1 Simulation Design

The simulation considers three exercises. The first studies measurement error and calibration. The second studies bounded allocation regret. The third studies whether AI-inferred utility improves prediction in a random utility model without implying recovery of latent utility.

The baseline data-generating process is:

$$u_i^* \sim N(0, 1) \quad (133)$$

$$\hat{u}_i^{\text{AI}} = u_i^* + \eta_i \quad (134)$$

$$u_i^H = u_i^* + \nu_i \quad (135)$$

$$Y_i = \alpha + \beta u_i^* + \gamma X_i + \varepsilon_i \quad (136)$$

where u_i^* is latent utility, $\hat{u}_i^{\text{AI}}$ is the AI-inferred utility proxy, u_i^H is the human validation benchmark, η_i is AI measurement error, and ν_i is benchmark error.

The simulation uses $\beta = 1$, $\gamma = 0.5$, and $\nu_i \sim N(0, 0.25^2)$. The calibrated proxy is constructed by estimating the discrepancy:

$$\Delta_i = \hat{u}_i^{\text{AI}} - u_i^H \quad (137)$$

on the validation sample and then correcting the AI proxy:

$$\tilde{u}_i^{\text{AI}} = \hat{u}_i^{\text{AI}} - \hat{\Delta}_i. \quad (138)$$

The baseline simulation uses $N = 4,000$, $R = 300$ Monte Carlo replications, and validation sample size $n_V = 250$. The simulations are illustrative rather than empirical: they do not assume that benchmark accuracy rates from existing LLM studies (such as the 85% accuracy rate in Park et al. or the 47.8% accuracy rate in ConsumerSimBench) are direct estimates of utility-scale error or bounds, but reference them only as external motivation that AI behavioral proxies can be informative but imperfect.

The following figures visualize the same Monte Carlo results reported in the tables. They

are included only to clarify the mechanisms of measurement error, calibration, coefficient attenuation, and bounded regret; they do not introduce additional empirical claims.

B.2 Measurement Error and Calibration Results

The raw AI proxy produces substantial coefficient bias. Under classical error, the estimated coefficient falls from the true value $\beta = 1$ to about 0.444, consistent with attenuation bias. Under non-classical error, the raw proxy remains substantially biased. Calibration sharply reduces MSE and restores the coefficient close to the oracle benchmark.

Table B1: Measurement Error and Calibration Results

Scenario	MSE Raw	MSE Calib.	MSE Reduc.	Oracle β	Raw β	Calib. β	Raw Bias	Calib. Bias
Classical error	1.000	0.450	55.0%	1.000	0.444	1.000	-0.556	-0.000
Non-classical error	1.039	0.194	81.4%	1.000	0.507	1.001	-0.493	0.001

Interpretation: The raw AI proxy produces substantial coefficient bias. Under classical error, the estimated coefficient falls from the true value $\beta = 1$ to about 0.444, consistent with attenuation bias. Under non-classical error, the raw proxy remains substantially biased. Calibration sharply reduces MSE and restores the coefficient close to the oracle benchmark.

Important wording: Note that calibration does not recover latent utility. It improves alignment with the latent benchmark in simulation and with the human validation benchmark in empirical settings.

B.3 Sensitivity to Validation Sample Size

Table B2: Sensitivity to Validation Sample Size

Val. Size n_V	MSE Raw	MSE Calib.	MSE Reduc.	Raw β	Calib. β	Calib. Bias
100	0.999	0.459	54.1%	0.445	1.011	0.011
250	0.998	0.451	54.8%	0.445	0.992	-0.008
500	1.001	0.448	55.3%	0.444	1.012	0.012

Interpretation: Even relatively small validation samples produce meaningful correction, but $n_V = 250$ is a reasonable minimum benchmark. Larger validation samples slightly stabilize calibration, but the main gain comes from having a representative validation sample.

B.4 Sensitivity to AI Measurement Error

Interpretation: The higher the AI measurement error, the more severe the raw attenuation bias. Calibration becomes more valuable precisely when the AI proxy is informative but noisy.

Table B3: Sensitivity to AI Measurement Error

σ_η	MSE Raw	MSE Calib.	MSE Reduc.	Raw β	Calib. β	Raw Bias	Calib. Bias
0.25	0.063	0.059	5.1%	0.926	1.001	-0.074	0.001
0.50	0.250	0.193	22.7%	0.760	0.998	-0.240	-0.002
1.00	0.998	0.448	55.1%	0.444	1.009	-0.556	0.009
1.50	2.247	0.599	73.4%	0.263	0.992	-0.737	-0.008

The simulation also shows that calibration becomes increasingly valuable as AI measurement error rises: MSE reduction increases from 5.1% when $\sigma_\eta = 0.25$ to 73.4% when $\sigma_\eta = 1.50$, indicating that weaker or noisier AI proxies require stronger validation and calibration.

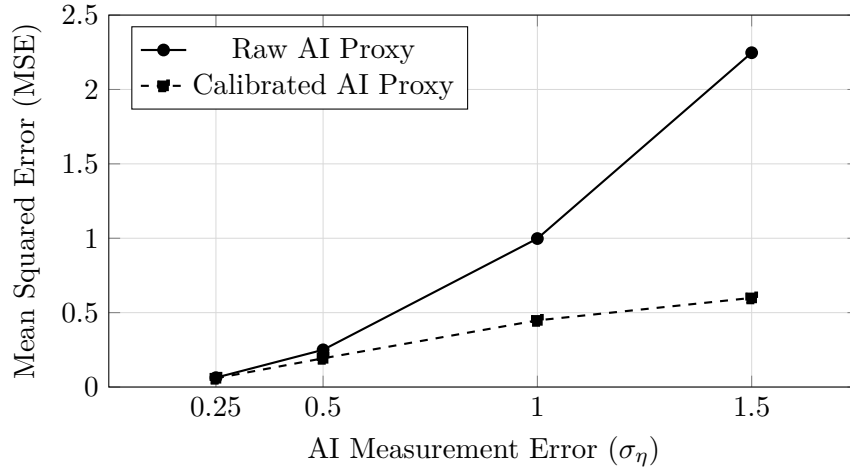


Figure B1: Calibration reduces mean squared error, with the largest gains occurring when AI measurement error is high.

B.5 Confidence Interval Coverage

Table B4: Confidence Interval Coverage

Scenario	Mean Calib. β	Mean Naive SE	Mean Boot. SE	Naive 95% Cov.	Boot. 95% Cov.
Classical error	1.016	0.045	0.093	61.7%	91.7%
Non-classical error	1.011	0.032	0.052	73.3%	91.7%

Interpretation: Calibration corrects the point estimate, but standard inference must account for the additional uncertainty introduced by estimating the calibration function. This supports the paper’s recommendation that AI-inferred variables require validation samples and bootstrap or adjusted standard errors.

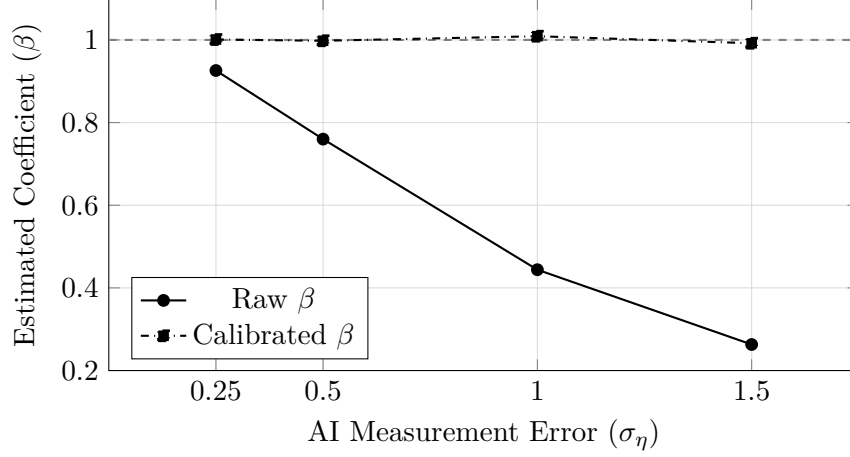


Figure B2: Raw AI-inferred utility produces attenuation as measurement error increases, while calibration restores the coefficient close to the oracle benchmark.

B.6 Bounded Allocation Regret Simulation

For each individual i , generate $J = 10$ alternatives with latent utilities:

$$u_{ij}^* \sim N(0, 1) \quad (139)$$

and calibrated AI utilities:

$$\tilde{u}_{ij}^{\text{AI}} = u_{ij}^* + \eta_{ij} \quad (140)$$

where:

$$\eta_{ij} \sim \text{Uniform}(-\epsilon, \epsilon). \quad (141)$$

Define the true optimal choice:

$$j^* = \arg \max_j u_{ij}^* \quad (142)$$

and the AI-based choice:

$$\hat{j} = \arg \max_j \tilde{u}_{ij}^{\text{AI}}. \quad (143)$$

The regret is:

$$\text{Regret}_i = u_{ij^*}^* - u_{i\hat{j}}^* \quad (144)$$

The theorem predicts:

$$\text{Regret}_i \leq 2\epsilon. \quad (145)$$

Interpretation: The numerical simulation confirms the theoretical regret bound. Even as AI proxy error increases, realized allocation regret remains below 2ϵ . This supports the paper's

Table B5: Bounded Allocation Regret Simulation

ϵ	Avg. Regret	95th Pct. Regret	Max. Regret	Bound 2ϵ	Violation Rate	Wrong Choice Rate
0.05	0.0007	0.0000	0.0954	0.10	0.0%	2.55%
0.10	0.0024	0.0000	0.1753	0.20	0.0%	4.89%
0.25	0.0158	0.1352	0.4789	0.50	0.0%	12.33%
0.50	0.0587	0.4021	0.9586	1.00	0.0%	23.37%
1.00	0.1986	0.9429	1.8822	2.00	0.0%	40.78%

claim that operational allocation may be useful even without perfect latent utility recovery. An additional implication is that a high wrong-choice rate does not necessarily imply large welfare loss: even when $\epsilon = 1.00$, the AI-based choice is wrong in 40.78% of cases, but the average regret remains only 0.1986 utility units and stays below the theoretical 2ϵ bound. This indicates that many wrong choices are close substitutes to the true optimum.

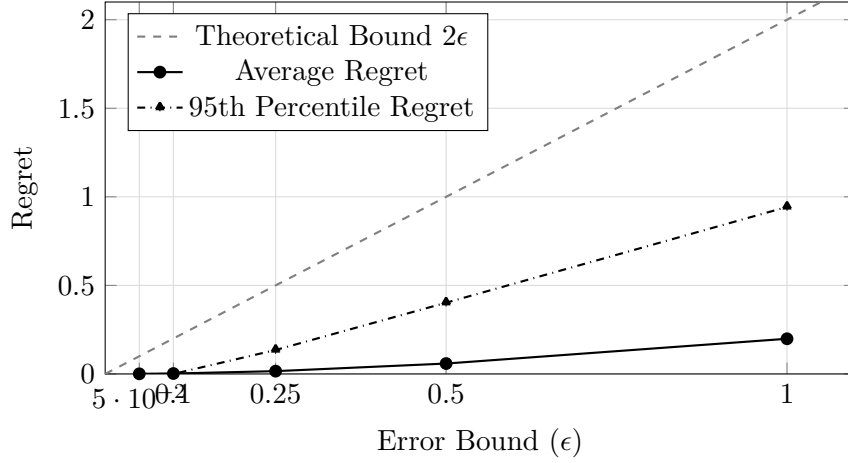


Figure B3: Simulated allocation regret remains below the theoretical upper bound 2ϵ , illustrating the bounded-regret result.

B.7 Random Utility Model Simulation

The final simulation embeds the calibrated AI proxy into a multinomial logit model:

$$U_{ij} = V_{ij} + \epsilon_{ij} \quad (146)$$

$$V_{ij} = X'_{ij}\beta - \lambda p_j + \rho \tilde{u}^{\text{AI}}_{ij} \quad (147)$$

The baseline model excludes $\tilde{u}^{\text{AI}}_{ij}$, while the AI-augmented model includes it. The purpose is not to show that AI recovers latent utility, but to test whether the AI proxy provides incremental predictive content.

The calibration slope is estimated as $\hat{b} = 0.6349$.

Table B6: Random Utility Model Simulation

Model	Train LL/Choice	Test LL/Choice	Test Acc.	β_X	λ_p	ρ_{AI}
Baseline (X, p only)	-1.4217	-1.4309	39.85%	0.6303	-0.9106	—
AI-augmented ($X, p, \hat{u}^{AI}$)	-1.3141	-1.3327	45.15%	0.6749	-0.9841	0.7316

Interpretation: Adding calibrated AI-inferred utility improves out-of-sample log-likelihood and choice accuracy. The positive estimate $\rho_{AI} = 0.7316$ indicates that the calibrated AI proxy is predictive in the simulation. However, this does not imply that the AI observes latent utility; it only shows that the proxy is informative for choice prediction after calibration.

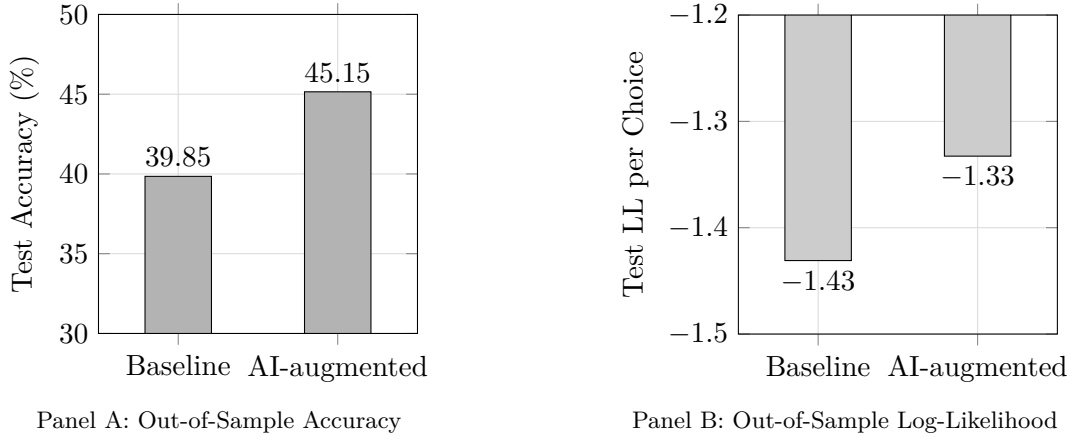


Figure B4: Adding calibrated AI-inferred utility improves out-of-sample prediction in the random utility model, without implying recovery of latent utility.

B.8 Summary of Simulation Evidence

The Monte Carlo exercises validate the three operational claims of Theorem 1. First, calibration reduces measurement error substantially, lowering MSE by 55.0% under classical measurement error and by 81.4% under non-classical measurement error (Table B1), supporting the calibration logic in Proposition 6. Second, realized allocation regret remains strictly below the theoretical upper bound 2ϵ across all simulated error magnitudes, with a violation rate of 0.0% in every case (Table B5), supporting Proposition 8. Third, the AI-augmented random utility model improves out-of-sample prediction accuracy from 39.85% to 45.15% (Table B6), supporting the claim that AI-inferred utility can contain incremental predictive information. These results should be interpreted as illustrative Monte Carlo evidence: they show that noisy AI utility proxies can improve measurement, calibration, and allocation under structured assumptions, but they do not imply recovery of latent utility or causal identification.

8 References

1. Acemoglu, D. (2024). *The Simple Macroeconomics of AI* (NBER Working Paper No. 32487). National Bureau of Economic Research.
2. Acemoglu, D., Lin, T., Ozdaglar, A., & Siderius, J. (2026). *How AI Aggregation Affects Knowledge* (NBER Working Paper No. 35036).
3. Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1), 67–77.
4. Allen, R. G. D. (1934). The nature of indifference curves. *The Review of Economic Studies*, 1(2), 110–121.
5. Arrow, K. J. (1951). *Social Choice and Individual Values*. Wiley.
6. Arrow, K. J. (1965). *Aspects of the Theory of Risk-Bearing*. Yrjö Jahnssonin Säätiö.
7. Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.
8. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
9. Debreu, G. (1954). Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision Processes* (pp. 159–165). Wiley.
10. Debreu, G. (1959). *Theory of Value*. Yale University Press.
11. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). *Algorithmic Collusion by Large Language Models* (arXiv:2404.00806).
12. Hayek, F. A. (1935). *Collective Economic Planning*. Routledge.
13. Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, 35(4), 519–530.
14. Hicks, J. R. (1939). *Value and Capital*. Oxford University Press.

15. Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents* (NBER Working Paper No. 31122).
16. Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, 17(67), 159–174.
17. Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of Local Average Treatment Effects. *Econometrica*, 62(2), 467–475.
18. Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary policy according to HANK. *American Economic Review*, 108(3), 697–743.
19. Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. *RAND Journal of Economics*, 52(3), 538–558.
20. Korinek, A. (2023). Generative AI for economic research. *Journal of Economic Literature*, 61(4), 1281–1317.
21. Korinek, A. (2025). *AI Agents for Economic Research* (NBER Working Paper No. 34202).
22. Lange, O. (1936). On the economic theory of socialism: Part one. *The Review of Economic Studies*, 4(1), 53–71.
23. Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). *Large Language Models: An Applied Econometric Framework* (NBER Working Paper No. 33344).
24. McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press.
25. Mises, Ludwig von. (1920). Economic calculation in the socialist commonwealth. *Archiv für Sozialwissenschaft und Sozialpolitik*.
26. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative AI. *Science*, 381(6654), 187–192.
27. Peng, S., et al. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot* (arXiv:2302.06590).
28. Peterson, A. J. (2024). *AI and the Problem of Knowledge Collapse* (arXiv:2404.03502).

29. Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32(1/2), 122–136.
30. Robbins, L. (1932). *An Essay on the Nature and Significance of Economic Science*. Macmillan.
31. Samuelson, P. A. (1938). A note on the pure theory of consumer’s behaviour. *Economica*, 5(17), 61–71.
32. Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, 15(60), 243–253.
33. Smets, F., & Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3), 586–606.
34. Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica*, 50(4), 945–973.
35. von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.
36. Yao, S., et al. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models* (arXiv:2210.03629).