Revealed Rationality:
Label-Free Regularization from Representation Theorems

By Isaiah Andrews[1]

Abstract

Representation theorems in decision theory establish that behavior satisfies certain axioms if and only if it can be rationalized by a well-defined objective. I argue that this "if and only if" structure provides a natural foundation for label-free regularization of large language models and other AI systems. Axiom compliance can be checked from the model's own responses to synthetic choice problems, with no external labels or human feedback, and the penalties are computable in polynomial time. I develop three instantiations: probabilistic coherence via de Finetti's theorem, preference rationality via Afriat's theorem, and subjective expected utility via Echenique and Saito (2015), each yielding a continuous penalty that equals zero if and only if behavior can be rationalized. Since coherence does not restrict what objective rationalizes behavior, these penalties complement rather than replace other training signals.

Keywords: Representation theorems, Rationality, Regularization, Large language models

1 Introduction

Large language models are increasingly used in contexts that require them to express beliefs, rank alternatives, and recommend actions. Recent work suggests both that the rationality of these responses varies substantially across models and that it is responsive to design choices. Tak et al. (2026), for instance, test whether large language models satisfy the von Neumann-Morgenstern axioms (completeness, transitivity, continuity, independence) and find widespread violations, but also find that models using additional reasoning tokens show substantially improved axiom compliance. Chadwick et al. (2025) observe that incoherent probability assessments and intransitive preference orderings in large language models are directly exploitable by constructing Dutch books (combinations of bets guaranteed to extract money from the model) and money pumps (sequences

[1]First version: February 21, 2026. Department of Economics, Massachusetts Institute of Technology, and NBER, iandrews@mit.edu. I thank Jiafeng Chen and Sendhil Mullainathan for useful discussions, and Claude Opus 4.6 for excellent research assistance.

of trades which extract value) respectively. Together, these findings illustrate both that rationality violations can be consequential and that they are amenable to intervention.

A broader diagnostic literature reinforces these observations. Chen et al. (2023) test whether GPT-3.5 choices from budget sets satisfy axioms from revealed preference theory, finding high but imperfect compliance. Hagendorff et al. (2023) show that earlier generations of LLMs depart from rational choice, but that these violations are smaller for later model generations. Mazeika et al. (2025) find that larger models largely satisfy expected utility axioms, but with sometimes unappealing emergent goals. Qiu et al. (2026) find that LLMs fail to correctly apply Bayes rule, but that performance is dramatically improved by training on the predictions of a Bayesian model. This literature reinforces that rationality violations are present and, in some cases, responsive to training, but also that coherence alone does not settle the question: a model can satisfy axioms of rationality while pursuing objectives that are misaligned.

Most of this work is diagnostic, documenting the presence or absence of particular rationality properties. Another possibility, however, is to use the implications of rationality directly in training, penalizing violations in the same way that other constraint-based penalties are used in machine learning. I argue that classic representation theorems in decision theory provide a natural foundation for such an approach. A representation theorem is a result of the form: behavior satisfies axioms $A_1, \ldots, A_k$ if and only if that behavior can be rationalized by some objective (e.g. a probability measure, a utility function, or both).[2] The forward direction says that any "rational" agent (in the sense specified by the theorem) satisfies the axioms, which is expected and consistent with the diagnostic literature discussed above. The theorems also cover the reverse direction, however, establishing that if a model's behavior satisfies the axioms, then there exists a well-defined objective that rationalizes it. This seems potentially powerful in the context of model training, since a training procedure need not specify what that objective is. It only needs to drive axiom violations toward zero, and the representation theorem guarantees that the resulting behavior can be rationalized.

Axiom compliance can be checked using only the model's own responses. One could

[2]The term "representation" here refers to the rationalization of behavior by a mathematical objective, not to the feature representations or embeddings discussed in machine learning.

for instance generate synthetic choice problems (betting scenarios, budget sets, portfolio allocations), query the model, and check whether the responses satisfy the relevant axioms. No external labels, ground truth outcomes, or human feedback is required. The combinatorial explosion that limits axiom-based testing for human subjects may be a feature in the LLM context: synthetic problems can be generated cheaply and in large numbers. Moreover, as I discuss below the relevant penalties can in many cases be computed in polynomial time.

It is important to emphasize that coherence is not sufficient for good behavior. A model satisfying all the axioms of subjective expected utility has a well-defined probability measure and utility function but both could be terrible, so other training signals remain essential. The role of axiom-based regularization is to push the model towards internal consistency: whatever the model does, it should do coherently.

I discuss three instantiations of this idea, in order of increasing richness. Section 2 considers probabilistic coherence through de Finetti's theorem: a model's probability assessments are consistent with a well-defined probability measure if and only if they cannot be Dutch-booked. Section 3 considers preference rationality through Afriat's theorem: a model's choices from budget sets are consistent with utility maximization if and only if they satisfy the Generalized Axiom of Revealed Preference. Section 4 considers the joint structure of beliefs and preferences under uncertainty through results of Echenique and Saito (2015): a model's portfolio choices are consistent with subjective expected utility maximization (with a concave utility) if and only if they satisfy the Strong Axiom of Revealed Subjective Expected Utility. Section 5 briefly discusses implementation. Section 6 situates the approach relative to RLHF, calibration, and other suggestions from the literature. Section 7 discusses limitations and concludes.

2 Probabilistic coherence: de Finetti

The simplest instance of the general principle concerns probabilistic beliefs. De Finetti's coherence theorem provides an "if and only if" characterization of when a set of probability assessments is consistent with a well-defined probability measure.

Setup Consider a finite collection of events $E_1, \ldots, E_n$ defined on a state space Ω. A *prevision assignment* (i.e., a set of probabilistic predictions) is a function p that assigns a value $p(E_i) \in [0, 1]$ to each event. In our context, these are the model's stated probability assessments when queried about each event.

Avoiding sure loss A bet on event E_i with stake $b_i \in \mathbb{R}$ yields a payoff of $b_i(\mathbf{1}_{E_i}(w) - p(E_i))$ to the bettor when the state is ω, where $\mathbf{1}_{E_i}(\omega)$ takes the value one when $\omega \in E_i$ and zero otherwise.[3] A *Dutch book* against the prevision assignment p is a collection of stakes $(b_1, \ldots, b_n)$ such that the agent's net payoff $-\sum_{i=1}^{n} b_i(\mathbf{1}_{E_i}(\omega) - p(E_i))$ is strictly negative in every state of the world ω. The existence of a Dutch book means that the agent's stated probabilities are exploitable: an adversary can construct a combination of bets that guarantees a profit regardless of which events occur. The prevision assignment p *avoids sure loss* if no Dutch book exists against it.

The representation theorem De Finetti's coherence theorem (de Finetti, 1937, 1974) states that the following are equivalent:

1. The prevision assignment p avoids sure loss.

2. The prevision assignment p can be extended to a finitely additive probability measure on the algebra generated by $E_1, \ldots, E_n$.

Condition (1) depends only on the agent's stated probabilities and can be checked without knowing which events actually occur. Condition (2) says that the probabilities can be rationalized by a coherent probability measure. The equivalence between (1) and (2) is the representation theorem. Note that finitely additive probability suffices since we consider only a finite collection of events.

The penalty The magnitude of the best Dutch book against a prevision assignment p provides a natural, continuous penalty. Let $\omega_1, \ldots, \omega_m$ denote the atoms of the algebra generated by $E_1, \ldots, E_n$. The bettor's profit in state ω_j from a collection of stakes

[3]The convention is that the AI agent takes the opposite side of the bet, so the agent's payoff from this bet is $-b_i(\mathbf{1}_{E_i}(\omega) - p(E_i))$.

$(b_1, \ldots, b_n)$ is

$$\Pi(\omega_j) = \sum_{i=1}^{n} b_i \big(\mathbf{1}_{E_i}(\omega_j) - p(E_i) \big).$$

The sure-loss magnitude is the optimal value of the linear program

$$L(p) = \max_{b \in \mathbb{R}^n} \min_{j=1,\ldots,m} \Pi(\omega_j), \tag{1}$$

subject to $\sum_i |b_i| \leq 1$. The normalization bounds the total size of all bets, so $L(p)$ measures the guaranteed profit per unit of total stake. By the representation theorem, $L(p) = 0$ if and only if p is coherent. When $L(p) > 0$, its value quantifies the exploitability of the model's stated probabilities. This is a linear program and is solvable in polynomial time (in the number of atoms and events).[4]

Translation to the LLM context To apply this in training, one may generate a partition of a sample space into atoms, along with a collection of events defined as unions of those atoms (e.g. that it rains tomorrow in Boston, but not New York, and there is at least one hour of sun during the day in Boston, and ...). The model is presented with descriptions of the events and asked to report a probability for each. The logical relationships among events (set inclusion, partitioning, complementation) are known by construction, so the full LP can be assembled from the model's responses without any external labels.

The constraints checked jointly by the Dutch book LP include finite additivity, monotonicity, and the law of total probability. When conditional events are included in the elicited algebra, coherence also enforces Bayes' rule. These are all implications of the probability axioms, and the LP checks them simultaneously.

An important design consideration is that the axiom checks test consistency across a collection of assessments attributed to a single agent. In the LLM context, it is natural to fix a role or persona for each batch of queries (e.g., "you are a forecaster assessing weather probabilities") and to test coherence within that role, rather than across different

[4]Garrabrant et al. (2016) use a related no-arbitrage criterion, requiring that no computable trading strategy earn unbounded profits, as the coherence condition for a logical probability assigner. Their setting (logical statements) differs from the empirical events considered here, but the underlying principle is the same: the absence of profitable exploitation characterizes coherence.

roles, since it is not clear that probability assessments should be consistent across distinct identities.

Existing evidence Zhu and Griffiths (2024) directly measure probability coherence violations in LLM outputs across several model families, finding systematic incoherence in probability judgments. Betz and Richardson (2023) demonstrate that self-training improves the probabilistic coherence of neural language models trained on synthetic data. Qiu et al. (2026) train models for Bayesian probabilistic reasoning via supervised fine-tuning on demonstrations from an external model, and show that this generates performance gains which generalize to new tasks. Zhu et al. (2025) train a variational autoencoder on LLM embeddings constrained to satisfy the complementary rule $P(A)+P(\neg A) = 1$, and show that the resulting internal model probabilities are both more coherent and better calibrated. These results demonstrate that probabilistic coherence can be improved by training; the present approach differs in using a continuous penalty grounded in the representation theorem rather than supervised examples or checks for a subset of coherence properties.

3 Preference rationality: Afriat

The second instantiation concerns preferences over bundles. Afriat's theorem provides a revealed preference test for utility maximization that is directly analogous to de Finetti's theorem for probabilistic coherence: behavior is rationalizable if and only if it satisfies an axiom stated on the observed choices.

Setup The data consist of T observations $\{(x_t, p_t)\}_{t=1}^{T}$, where $x_t \in \mathbb{R}^K_+$ is the bundle chosen at prices $p_t \in \mathbb{R}^K_{++}$, and corresponds to income $w_t = p_t \cdot x_t$.

GARP The choices $\{(p_t, x_t)\}_{t=1}^{T}$ reveal information about preferences. If x_t was chosen when x_s was affordable ($p_t \cdot x_s \leq p_t \cdot x_t$), then x_t is *directly revealed preferred* to x_s, written $x_t \; R^D \; x_s$. The *revealed preference* relation R is the transitive closure of R^D - for instance, if $x_t R^D x_s$ and $x_s R^D x_r$ then $x_t R x_r$.

The *Generalized Axiom of Revealed Preference* (GARP) requires: if $x_t \; R \; x_s$, then $p_s \cdot x_t \geq w_s$. That is, if x_t is revealed preferred to x_s (directly or through a chain), then x_t must not have been strictly inside the budget set at which x_s was chosen.

The representation theorem Afriat's theorem (Afriat, 1967; Varian, 1982) states that the following are equivalent:

1. The data $\{(p_t, x_t)\}_{t=1}^T$ satisfy GARP.
2. There exist numbers $\{u_t, \lambda_t\}_{t=1}^T$ with $\lambda_t > 0$ satisfying the *Afriat inequalities*: $u_s - u_t \leq \lambda_t p_t \cdot (x_s - x_t)$ for all s, t.
3. There exists a continuous, monotone, concave utility function $U : \mathbb{R}_+^K \to \mathbb{R}$ such that

$$x_t \in \arg\max_{x : p_t \cdot x \leq w_t} U(x) \text{ for all } t.$$

The equivalence between (1) and (3) is the representation theorem: choices are consistent with utility maximization if and only if they satisfy GARP. If any locally nonsatiated (i.e., more is at least weakly preferred to less) utility function rationalizes the data, a continuous, monotone, concave one does as well (Afriat, 1967). The Afriat inequalities in (2) are a finite system of linear inequalities providing the constructive check for whether a finite dataset is consistent with GARP. Note that the rationalizing utility is not unique: many different utility functions may be consistent with the same finite dataset. The representation theorem guarantees existence, not identification.

The penalty: CCEI GARP is a binary condition: it either holds or it fails. For regularization, a continuous measure of departure is needed. The *Critical Cost Efficiency Index* (CCEI), introduced by Afriat (1973), provides a natural such measure.

Define *e-GARP* as the relaxation in which budget constraints are tightened by a factor $e \in [0, 1]$: x_t is directly revealed preferred to x_s under e-GARP only if $p_t \cdot x_s \leq e \cdot p_t \cdot x_t$. The CCEI is the largest e such that the data satisfy e-GARP:

$$\text{CCEI} = \sup\{e \in [0, 1] : \{(p_t, x_t)\}_{t=1}^T \text{ satisfies } e\text{-GARP}\}.$$

When CCEI $= 1$, the data satisfy GARP exactly; the penalty $1 - \text{CCEI}$ measures the fraction by which budgets must be reduced to rationalize the data. The CCEI is computable by binary search over e, with each step requiring a GARP check. The penalty is continuous, piecewise linear, bounded in $[0, 1]$, and equals zero if and only if the data are rationalizable.[5]

Translation to the LLM context To use this approach in model training, one could present the model with a role and a budget constraint (prices and income), and ask it to allocate across goods. One could then vary prices and income, compute the CCEI of the resulting choices, and penalize $1 - \text{CCEI}$. The decision context can be varied across batches: one might involve grocery purchases, another retirement portfolio allocation, another time allocation across tasks. GARP requires consistency within each context, and testing across many contexts strengthens the regularization. As in the de Finetti case, it seems natural to fix a role or persona within each batch and test consistency within that role.

Existing evidence Chen et al. (2023) apply the GARP framework to test whether GPT models make rational choices. They present GPT-3.5 with budget allocation tasks across four domains (risk, time, social preferences, and food) and compute the CCEI. The average CCEI values exceed 0.997 across all domains, outperforming human subjects in comparable experimental designs. Importantly, however, these tests use two-good settings with a relatively limited number budget sets. Thus, while Chen et al. (2023) demonstrate that their tests have sufficient power for their purposes, GARP-based regularization may have more bite in higher dimensions and with more observations, where it is much easier to gather data from LLMs than from human subjects. Wen (2025) finds that prompt-based role specialization (assigning domain-specific personas such as "biotechnology expert" or "economist") substantially reduces GARP compliance, with specialized agents showing lower CCEI values than the baseline. Seror (2024) extend

[5]The money pump indices of Echenique et al. (2011) provide a complementary "exploitability" interpretation analogous to the Dutch book. Smeulders et al. (2013) show that versions of this index discussed in Echenique et al. (2011) are NP-hard to compute, but propose alternative money-pump indices that are computable in polynomial time and so may also be appealing for training purposes. See also Echenique (2021), who highlights important limitations in the interpretability of the CCEI.

GARP testing to moral preference domains, finding variation in choices across models.

4 Decisions under uncertainty: Echenique-Saito

The preceding sections address beliefs and preferences separately. However, a model could have coherent probabilities and rational preferences but still fail to integrate them. The third instantiation of the general principle addresses this by jointly constraining beliefs and preferences under uncertainty.

Setup Let $S = \{1, \ldots, \bar{s}\}$ be a finite set of states. A *monetary act* is a vector $x \in \mathbb{R}^S_+$ specifying a payoff $x(s)$ in each state s. The data consist of T observations $\{(x_t, p_t)\}_{t=1}^T$, where $x_t \in \mathbb{R}^S_+$ is the portfolio chosen at prices $p_t \in \mathbb{R}^S_{++}$ for state-contingent claims (securities that pay one unit in a designated state and zero otherwise) with income $w_t = p_t \cdot x_t$.[6]

SARSEU Echenique and Saito (2015) introduce the following condition.

Definition 1 (SARSEU)**.** *A data set $\{(x_t, p_t)\}_{t=1}^T$ satisfies the* Strong Axiom of Revealed Subjective Expected Utility *(SARSEU) if, for any sequence of pairs $(x_{t_i}(s_i), x_{t'_i}(s'_i))_{i=1}^n$ such that*

(i) $x_{t_i}(s_i) > x_{t'_i}(s'_i)$ for all i,

(ii) each state s appears as s_i (on the left) the same number of times as it appears as s'_i (on the right),

(iii) each observation t appears as t_i (on the left) the same number of times as it appears as t'_i (on the right),

we have

$$\prod_{i=1}^{n} \frac{p_{t_i}(s_i)}{p_{t'_i}(s'_i)} \leq 1. \tag{2}$$

See Echenique and Saito (2015) for discussion of these conditions.

[6]The restriction to monetary payoffs (one good per state) is a limitation relative to the Afriat setting, which handles bundles of multiple goods. This is, however, a common setting for considering decision under uncertainty, particularly in asset pricing. In addition, the extension to state-dependent utility discussed below considerably generalizes the result.

The representation theorem Echenique and Saito (2015) show the following are equivalent

1. The data $\{(p_t, x_t)\}_{t=1}^{T}$ satisfy SARSEU.

2. There exist a full-support probability measure $\mu \in \Delta_{++}(S)$ and a concave, strictly increasing function $u : \mathbb{R}_+ \to \mathbb{R}$ such that

$$x_t \in \arg\max_{x : p_t \cdot x \leq w_t} \sum_{s \in S} \mu(s) u(x(s)) \text{ for all } t.$$

Condition (2) requires that the agent has a single probability distribution over states (beliefs) and a single utility function over monetary payoffs (preferences), and chooses x to maximize expected utility. Savage (1954) provided a set of axioms for preferences over "acts" mapping states to "consequences" (potentially more general than simple monetary payoffs) and proved that, under these axioms, preferences admit an SEU representation in the sense of (2). Savage's framework, however, restricts the preference ordering over all acts, which is not what one observes in finite data. Echenique and Saito (2015) provide the operational analog for the problem with monetary payoffs: the SARSEU restriction in (1) is a revealed-preference characterization of SEU that applies to finite datasets of choices at prices, exactly as Afriat (1967) operationalizes utility theory for finite budget-set data.

As in the de Finetti and Afriat cases, the condition is necessary and sufficient: the axiom holds if and only if behavior is consistent with subjective expected utility maximization. Echenique and Saito (2015) show in the proof of their Proposition 2 that SARSEU can be tested by checking feasibility of a linear program.

The penalty: SEU efficiency index By direct analogy with the CCEI of Section 3, one may define an efficiency index for the SEU problem. In particular, the linear program considered in Echenique and Saito (2015)'s Proposition 2 effectively checks for the existence of cycles violating (2). To define an efficiency index $e \in [0, 1]$, we may consider a modified version of (2) which scales the price ratio by e, and let e^*_{SEU} denote the largest

e such that

$$\prod_{i=1}^{n} e \cdot \frac{p_{t_i}(s_i)}{p_{t'_i}(s'_i)} \leq 1.$$

for all $(x_{t_i}(s_i), x_{t'_i}(s'_i))_{i=1}^{n}$ satisfying the SARSEU conditions. $e^*_{\text{SEU}} = 1$ if and only if the SARSEU conditions hold, and checking whether a given e satisfies these conditions simply requires running a modified version of Echenique and Saito (2015)'s linear program. Computation of e^*_{SEU} proceeds by binary search over e: by construction, making e smaller only makes it easier to satisfy the axiom, so the set of passing values is an interval.

Translation to the LLM context The elicitation procedure is similar to that of Section 3. Present the model with a set of states of the world. Offer state-contingent claims at given prices and budgets. The model allocates a budget across states. Multiple rounds with different price vectors yield the dataset $\{(x_t, p_t)\}_{t=1}^{T}$. As in the preceding sections, the model should be assigned a fixed role within each batch, since the approach tests consistency of choices attributed to a single decision-maker.

Existing evidence No direct tests of SEU rationality (SARSEU or related conditions) on LLM-generated data appear to exist. Mazeika et al. (2025) test expected utility axioms such as independence and monotonicity, finding high compliance in large models, but do not test the joint belief-preference structure that SARSEU captures. Tak et al. (2026) likewise test model adherence with the expected utility axioms, though they further show that models exhibit ambiguity aversion in settings with unknown probabilities. Yamin et al. (2026) elicit probability estimates from LLMs in decision tasks and test whether these beliefs satisfy properties required of a rational decision-maker, including decision-sufficiency (whether stated beliefs fully explain the model's choices) and monotonicity of choice probabilities. They find what they characterize as violations, suggesting that LLM-reported probabilities do not form a fully coherent basis for the decisions the models actually make. This is suggestive of the kind of joint incoherence that SARSEU is designed to detect, though their framework (random utility models with elicited beliefs) differs from the revealed-preference approach taken here.

Extensions: alternative axiom systems Full SEU may be too strong a rationality standard. In particular, SEU rules out state-dependent utility $u_s(\cdot)$, as could arise if the agent anticipates higher consumption needs in some states than others. It also rules out ambiguity aversion, the tendency to prefer known risks over unknown ones (Knight, 1921), which some might view as desirable for a system facing genuine uncertainty.

Fortunately, existing results suggest paths forward in both cases. Echenique and Saito (2015) provide a generalization of their results to settings with state-dependent utility, which again gives if and only if conditions for a finite dataset to be compatible with subjective utility maximization with utility $u_s(\cdot)$. Similarly, representation theorems are available for weaker axiom systems which allow ambiguity aversion. Gilboa and Schmeidler (1989) weaken Savage's axioms, obtaining a representation by maxmin expected utility: the agent maximizes expected utility under the worst-case probability measure from a convex set of priors. Maccheroni et al. (2006) characterize variational preferences, a broader class that includes maxmin EU as a special case. Both models are testable via the GRID framework of Polisson et al. (2020), which provides another systematic procedure for testing classes of decision models using revealed-preference data.

The Anscombe-Aumann framework (Anscombe and Aumann, 1963) provides an alternative axiomatization that naturally extends to consequences richer than monetary payoffs (for example, bundles of goods). In this framework, acts map states to lotteries over consequences. Extending the present approach to this setting is a natural direction for future work.

The choice of axiom system is a substantive decision: different systems encode different notions of rationality and determine what behavior the regularization permits.

5 Implementation considerations

The suggested procedure The procedure for each of the three instantiations consists of three steps: (1) generate synthetic choice problems (collections of events for de Finetti, budget sets for Afriat, portfolio allocation problems for Echenique-Saito); (2) elicit responses from the model; (3) compute the penalty (the Dutch book magnitude $L(p)$, or $1 - \text{CCEI}$, or $1 - e^*_{\text{SEU}}$). The resulting penalty can then be used in the same way as other

constraint-based penalties in machine learning.

Practical considerations The choice problems used for axiom testing need not be sampled in any particular way. An adversarial approach to problem generation seems potentially appealing, with a second model proposing new choice problems based on previous performance (including the output from previous penalty computations).

An important practical concern is that the same formal choice problem, described in different words, or even the same words in a different order, may elicit different responses from the model. This is a violation of a foundational requirement: all three frameworks demand that the agent has well-defined preferences or beliefs over the underlying formal objects. In a controlled training setup, this is directly testable. Generate the formal problem first, then produce multiple linguistic descriptions of it and check that the model's responses are consistent across descriptions. If not, this "counts" as an axiom violation, and will be penalized by the approaches suggested above.

Computational costs All three penalty computations are polynomial-time: the Dutch book LP is polynomial in the number of atoms and events; while the GARP and SARSEU checks both involve solving linear programs, with dimensions depending on the number of choices (and states, in the case of SARSEU), together with bisection on $[0, 1]$. The penalties are continuous and piecewise linear in the model's numerical outputs.

6 Relation to existing approaches

The approach proposed here complements, rather than replaces, existing methods.

RLHF and preference structure. Current training methods including reinforcement learning from human feedback (RLHF; e.g. Ouyang et al., 2022) and direct preference optimization (DPO; e.g Rafailov et al., 2023) extract training signals from human feedback. Ge et al. (2024) provide an axiomatic critique of RLHF and propose an alternative, axiomatically grounded approach. All of these methods differ fundamentally from the focus of the present paper in that their focus is on generating reward signals for model training, rather than enforcing coherence on the model outputs. Since coherence of model outputs

is in important respects value-neutral, the approach discussed in the present paper is not a substitute for such reward signals

Calibration and proper scoring rules. Calibration methods (Guo et al., 2017) check whether predicted probabilities match empirical frequencies, requiring ground truth outcomes. De Finetti coherence checks internal consistency without ground truth. The two are complementary: a model can be well-calibrated on many events but incoherent, or coherent but poorly calibrated. Similarly, a strictly proper scoring rule incentivizes truthful reporting of individual probability estimates (Gneiting and Raftery, 2007). But proper scoring rules operate event by event and do not enforce consistency across multiple related assessments. Like calibration, they also require external outcomes data and are complementary to the coherence approach.

The diagnostic and correction literature. Many authors including Betz and Richardson (2023), Chen et al. (2023), Hagendorff et al. (2023), Zhu and Griffiths (2024), Chadwick et al. (2025), Wen (2025), and Qiu et al. (2026) have documented LLM rationality violations. As discussed above, some of these papers also discuss how such violations may be reduced, for instance through additional training steps as in Betz and Richardson (2023) and Qiu et al. (2026), or through post-processing of model outputs as in Chadwick et al. (2025). Relative to these approaches, the approach in this note is motivated by the "if and only if" guarantees provided by the representation theorems, which imply that if particular penalty terms could be driven to zero then there would necessarily exist a prior, utility function, or both which fully explain model behavior.

Utility engineering. The utility engineering agenda of Mazeika et al. (2025) concerns measuring and controlling the objectives implicit in LLM behavior, motivated by the finding that larger models develop increasingly coherent value systems, but these emergent objectives may be misaligned. This agenda appears highly complementary to the approach I suggest: utility engineering appears most applicable in settings where the model acts coherently enough for a rationalizing objective to exist, which is what the axiom-based regularization proposed here is designed to ensure. Conditional on passing the rationality checks proposed here, one could go further and examine the set of priors

and/or utility functions consistent with observed choices, which are natural inputs for a utility engineering exercise.

7 Limitations and conclusion

It is important to be clear about some limitations of the approach proposed here.

1. Coherence is not enough. A model that satisfies all the axioms of subjective expected utility has a well-defined probability measure and utility function, but these could be arbitrary.[7] The role of the penalties proposed here is regularization, not a standalone objective.

2. The choice of axiom system matters. Full subjective expected utility rules out ambiguity aversion and other departures that may be reasonable or desirable. Weaker axiom systems (maxmin EU, variational preferences) impose weaker rationality conditions and permit a broader range of behavior. The choice of axiom system is a substantive modeling decision.

3. The joint belief-preference test of Section 4 applies to monetary payoffs (one good per state). Extending to richer consequence spaces, for example bundles of goods via the Anscombe-Aumann framework, is a potentially useful direction for future work.

Classic results from decision theory characterize, in precise and testable terms, what it means for behavior to be rational. These results take the form of axioms checkable from behavior alone, with representation theorems guaranteeing that compliance implies the existence of a rationalizing objective. This structure appears well-suited to LLM training, where behavior can be observed at scale and tested systematically. The connection seems potentially fruitful in both directions: decision theory provides a natural collection of formal tools for regularization, while the AI model training context provides a potentially important setting for decision theorists interested in revisiting and extending this toolkit.

[7] More broadly, some authors (e.g. Zhi-Xuan et al., 2025) dispute whether rational choice frameworks of the sort presumed here are an appropriate foundation for LLM behavior in the first place.

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77.

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 14(2):460–472.

Anscombe, F. J. and Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, 34(1):199–205.

Betz, G. and Richardson, K. (2023). Probabilistic coherence, logical consistency, and bayesian learning: Neural language models as epistemic agents. *PLOS ONE*, 18(2):e0281372.

Chadwick, A., Kahng, A., and Kipper, J. (2025). Dutch books and money pumps: Rectifying vulnerabilities in LLMs through rationality. In *Proceedings of the 5th International Conference on Human and Artificial Rationality (HAR)*, Paris, France.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51).

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68.

de Finetti, B. (1974). *Theory of Probability, volume 1.* John Wiley & Sons, New York.

Echenique, F. (2021). On the meaning of the critical cost efficiency index on the meaning of the critical cost efficiency index. *arXiv preprint arXiv:2109.06354*.

Echenique, F., Lee, S., and Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6):1201–1223.

Echenique, F. and Saito, K. (2015). Savage in the market. *Econometrica*, 83(4):1467–1495.

Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N., and Taylor, J. (2016). Logical induction. *arXiv preprint arXiv:1609.03543*.

Ge, L., Halpern, D., Micha, E., Procaccia, A. D., Shapira, I., Vorobeychik, Y., and Wu, J. (2024). Axioms for AI alignment from human feedback. In *Advances in Neural Information Processing Systems*, volume 38.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Hagendorff, T., Fabi, S., and Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3:833–838.

Knight, F. H. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston.

Maccheroni, F., Marinacci, M., and Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498.

Mazeika, M. et al. (2025). Utility engineering: Analyzing and controlling emergent value systems in AIs. *arXiv preprint arXiv:2502.08640*.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 36.

Polisson, M., Quah, J. K.-H., and Renou, L. (2020). Revealed preferences over risk and uncertainty. *American Economic Review*, 110(6):1782–1820.

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T., and van Steenkiste, S. (2026). Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 37.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons, New York.

Seror, A. (2024). The moral mind(s) of large language models. *arXiv preprint arXiv:2412.04476*.

Smeulders, B., Cherchye, L., De Rock, B., and Spieksma, F. C. (2013). The money pump as a measure of revealed preference violations: A comment. *Journal of Political Economy*, 121(6):1248–1258.

Tak, A. N., Banayeeanzade, A., Bolourani, A., Bahrani, F., Chaubey, A., Karimireddy, S. P., Schwarz, N., and Gratch, J. (2026). Sparks of rationality: Do reasoning LLMs align with human judgment and choice? *arXiv preprint arXiv:2601.22329*.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–973.

Wen, S. (2025). Economic rationality under specialization: Evidence of decision bias in AI agents. *arXiv preprint arXiv:2501.18190*.

Yamin, K., Tang, J., Cortes-Gomez, S., Sharma, A., Horvitz, E., and Wilder, B. (2026). Do LLMs act like rational agents? Measuring belief coherence in probabilistic decision making. *arXiv preprint arXiv:2602.06286*.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. (2025). Beyond preferences in ai alignment. *Philisophical Studies*, 182:1813–1863.

Zhu, J.-Q. and Griffiths, T. L. (2024). Incoherent probability judgments in large language models. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*.

Zhu, J.-Q., Yan, H., and Griffiths, T. L. (2025). Recovering event probabilities from large language model embeddings via axiomatic constraints. *arXiv preprint arXiv:2505.07883*.