This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

Interpretable Deep Image Classification using Rationally Inattentive Utility Maximization

Kunal Pattanayak, *Student Member, IEEE*, Vikram Krishnamurthy, *Fellow, IEEE* and Adit Jain

***Abstract*—Can deep convolutional neural networks (CNNs) for image classification be interpreted as utility maximizers with information costs? By performing set-valued system identification for Bayesian decision systems, we demonstrate that deep CNNs behave equivalently (in terms of necessary and sufficient conditions) to rationally inattentive Bayesian utility maximizers, a generative model used extensively in economics for human decision-making. Our claim is based on approximately 500 numerical experiments on 5 widely used neural network architectures. The parameters of the resulting interpretable model are computed efficiently via convex feasibility algorithms. As a practical application, we also illustrate how the reconstructed interpretable model can predict the classification performance of deep CNNs with high accuracy. The theoretical foundation of our approach lies in Bayesian revealed preference studied in micro-economics. All our results are on GitHub and completely reproducible.**

***Index Terms*—Interpretable Machine Learning, Bayesian Revealed preference, Rational Inattention, Deep Neural Networks, Image Classification**

I. Introduction

This paper studies interpretable models for deep image classification. We propose a set-valued system identification approach to explain deep image classification. We show that image classification using deep Convolutional Neural Networks (CNNs) can be interpreted as a constrained Bayesian utility maximization problem where the observation likelihood is optimized, namely, maximize the expected utility subject to a cost constraint on the chosen observation likelihood. Such rationally inattentive Bayesian utility maximization models have recently been used to explain human decision-making in microeconomics.

In micro- and behavioral economics[1] a fundamental question relating to human decision-making is: *How to model attention spans in humans (agents)?* The area of rational inattention [1], [2], pioneered by Nobel laureate Christopher Sims, models human attention in information-theoretic terms. The key hypothesis is that agents are "boundedly rational"- their perception of the environment is modeled as a Shannon capacity limited channel. In simple terms, rational inattention assigns a mutual information cost for human attention spans.

Building on the rational inattention model, the next key concept is that of a Bayesian agent with rational inattention that maximizes its expected utility. Such models are studied extensively in micro-economics [3]–[5]. The intuition is this: more attentive decisions yield a higher expected utility at the expense of a larger attention cost. Hence, the Bayesian agent optimally trades off between minimizing its sensing cost and maximizing its expected utility. An important question is: *How to test for rationally inattentive utility maximization given the decisions of a Bayesian agent?* In the last decade, necessary and sufficient conditions have been developed in the area of Bayesian revealed preference [6], [7] to test if the decisions of a Bayesian agent are consistent with rationally inattentive utility maximization. In this paper, we use the necessary and sufficient conditions of [6], [7] to construct interpretable models for deep classification.

This non-parametric data-driven approach embeds the image classification task as a Bayesian utility maximization problem constrained by an information acquisition cost. We construct set-valued estimates of utility functions and information acquisition costs that rationalize deep image classification. In a signal processing context, the information cost, often referred to as the rational inattention cost in the literature, is analogous to the sensing cost incurred by a radar in controlled sensing [8], [9]. This approach to deep image classification can be viewed as an *inverse optimization* problem. Recently, neural networks have been used successfully to solve inverse problems in imaging [10]–[13]. However, to the best of our knowledge, an economics-based inverse optimization analysis of deep neural networks has not been explored in the literature.

Intuition. Rationally Inattentive Interpretable Deep Image Classification. From a deep learning perspective, a supervised classification model is optimized to minimize the misclassification loss between the true image labels (state) and the predicted image labels. The actions (predicted labels) of a trained neural network can be viewed as a black-box function evaluation of a perceived (noisy) version of the true state of the image. In this paper, we approximate the trained neural network's prediction model by a constrained Bayesian utility maximization model; see Fig. 1 for an illustration. In other words, the interpretable deep image classification approach in this paper can be interpreted as the system identification of the trained neural network assumed to be a rationally inattentive agent. The goal is to reconstruct feasible utilities and costs from the prediction behavior of the trained neural network, aggregated over several training parameters. The set-valued solutions for the convex feasibility tests outlined in Algorithms 1 and 4 yield utility functions and sensing costs that explain the trained neural networks' prediction behavior. Theorem 2 yields the sparsest estimate of the feasible variables

V. Krishnamurthy, K. Pattanayak and A. Jain are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853 USA. e-mail: vikramk@cornell.edu, kp87@cornell.edu, aj457@cornell.edu.

[1]Micro-economics models the interaction of individual agents pursuing their private interests. Behavioral economics models human decision-making in terms of subjective probabilities via prospect theory and framing. In the rest of this paper, we will use the term 'agent' to denote a Bayesian decision-maker.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

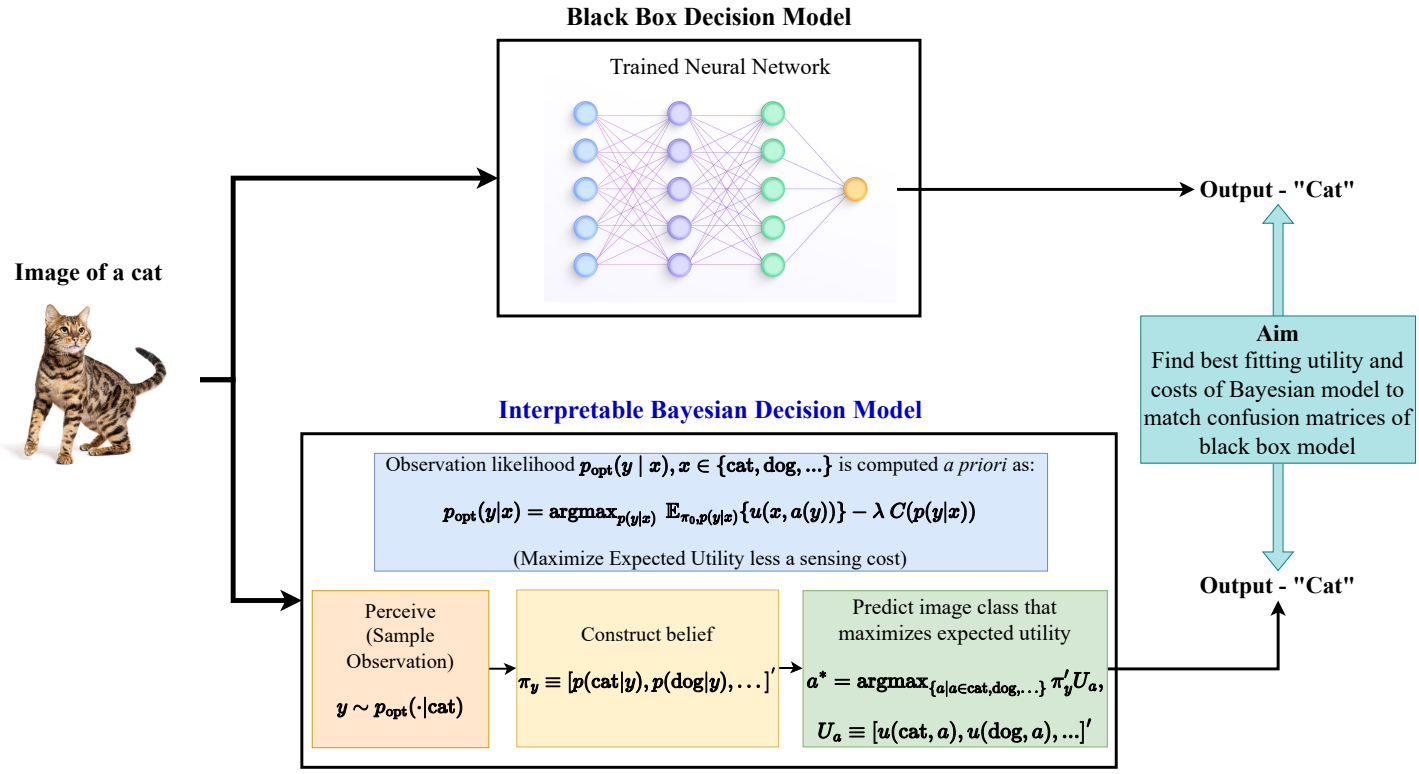


Fig. 1: Schematic for this paper's interpretable deep image classification approach. The neural network (top) is a black box that takes as input an image (for example, a cat image) and produces as output a predicted image label (for example, cat, dog, ship, etc.). This paper approximates the black box with a constrained Bayesian utility maximizer, with a constraint on the sensing/perception cost. The aim is to reconstruct parameters of the constrained Bayesian utility maximization model that best explain the prediction behavior of the trained neural network.

computed from Algorithms 1 and 4, while Definition 2 adopts a 'max-margin' approach to compute (a) goodness-of-fit of, and (b) robust utility and cost estimate for the rational inattention model that *rationalizes* the neural networks' behavior. The utility values computed from this interpretability approach to understand neural networks' decision-making process provide insights into how they prioritize classifying different classes under varying training parameters.

This paper uses a *data-driven* micro-economics based system identification approach for interpretable deep classification. The key ideas stem from Bayesian revealed preference [6], [7]. Bayesian revealed preference is a set-valued system identification algorithm for argmax non-linearity (in signal processing terms) that describes a Bayesian decision maker. Bayesian revealed preference is a *post-hoc* analysis of agent decisions. It constructs a generative[2] explanatory model for the agent decisions, parameterized by utility functions and an information acquisition cost. As a practical application, the interpretable model can also be used to predict the classification accuracy of the neural network trained on arbitrary training parameters; we discuss this in more detail in Sec. III-B. Our approach draws important parallels between human decision making and deep neural networks; namely that deep neural networks satisfy economics based rationality.

Why set-valued estimates of utility?: The aim of interpretable deep image classification is to construct feasible utility functions and information costs that rationalize neural network image label predictions over a finite set of training parameters. Estimating a utility function is an ill-posed problem (in the sense of Hadamard) since any non-negative increasing function of the utility is also a valid utility. From a statistical signal-processing perspective, a point-valued estimate is not useful for rationalizing a Bayesian decision maker's actions: (i) every point in the reconstructed set of feasible utilities and costs explains the actions equally well; hence, the problem is ill-posed, and (ii) a least squares estimate of the decision maker's utility function and information cost does not rationalize its actions. Bayesian revealed preference reconstructs a *set* of feasible utility functions and information acquisition costs that rationalize a Bayesian decision maker's actions in a finite number of environments. Every element in the feasible set explains the deep CNN decisions equally well. In Bayesian revealed preference, the utility functions are indexed by the environment; the information cost is invariant across environments. The computed utility function induces a preference ordering on the set of image classes. That is, how much a deep CNN prioritizes accurate classification over an inaccurate classification. The information acquisition cost abstracts the

[2]A generative model is image-independent, and hence provides a global explanation for deep image classification. In contrast, local approximation models for deep image classification are image-specific; they approximate model decisions via tractable functionals in a δ-neighborhood of every input.

penalty incurred by the deep CNN to 'learn' an accurate latent feature representation and can be interpreted as the training cost to achieve a desired accuracy of image classification.

A. Related Works

Since we study interpretable deep learning using behavioral and micro-economics, we briefly discuss related works in these areas.

Bayesian revealed preference and Rational inattention. Estimating utility functions given a finite sequence of decisions and budget constraints is the central theme of revealed preference in micro-economics. The seminal work of [14], [15] (see also [16]) give necessary and sufficient conditions for the existence of a utility function that rationalizes a finite time series of consumption bundles of a decision-maker. Rationally inattentive models for Bayesian decision making have been studied extensively in [3]–[5]. In the last decade, the area of Bayesian revealed preference [6], [7] develops necessary and sufficient conditions to test for rationally inattentive Bayesian utility maximization.

Interpretable ML. Providing transparent models for de-obfuscating 'black-box' ML algorithms under the area of interpretable machine learning is a subject of extensive research [17]–[19]. Interpretable machine learning is defined in [20] as "the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data".

Since the literature is enormous, we only discuss a subset of works pertaining to interpretability of deep neural networks for image classification [21], [22]. One prominent approach, namely, saliency maps, reconstructs the most preferred or typical image pertaining to each image class the deep neural network has learned [23], [24]. Related work includes creating hierarchical models for determining the importance of image features that determine its label [25]. This feature importance is encoded in this paper into the utility function that parametrizes our interpretable model. Another approach seeks to provide local approximations to the trained model, local w.r.t the input image [26], [27]. In contrast, our generative interpretable model provides a global black-box approximation for deep image classification. A third approach approximates the decisions of the deep neural networks by a linear function of simplified individual image features [27]–[30]. In contrast, our interpretable model fits a stochastic non-linear map that relates the true and predicted image labels. The parameters of the map are obtained by solving a convex feasibility problem parameterized by the deep CNN decisions. Finally, deep neural networks have also been modeled by Bayesian inference frameworks using probabilistic graphical methods [31].

To the best of our knowledge, an economics based approach for the post-hoc analysis of deep neural networks has not been explored in literature. However, we note that behavioral economics based interpretable models have been applied to domains outside interpretable machine learning, for example, in online finance platforms for efficient advertising [32], [33], training neural networks [34] and more recently in YouTube to rationalize user commenting behavior [35]. Finally, due to our recent equivalence result [36], our behavioral economics approach to interpretable deep image classification can be related to classical revealed preference methods [14], [15] in microeconomics.

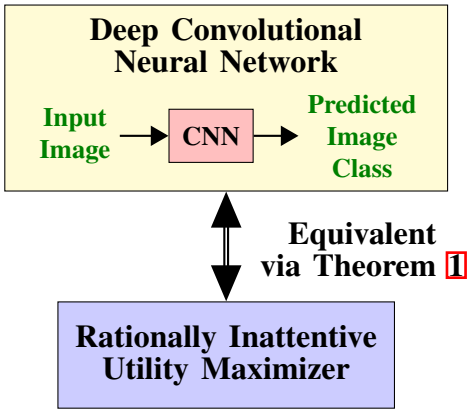


Fig. 2: Schematic illustration of rationally inattentive Bayesian utility maximization based interpretable image classification by deep CNNs. Theorem 1 establishes equivalence between the image classification behavior of a deep CNN and the decisions of a rationally inattentive maximizer. Hence, the deep CNN's image classification behavior can be parsimoniously represented by a utility function and an information acquisition cost.

B. Summary of Results.

The question we address is: *Can the decisions of deep CNNs in image classification be explained by a rationally inattentive Bayesian utility maximizer?*

The key results in this paper are:

1. We show that the image classification decisions of deep CNNs satisfy the necessary and sufficient conditions for rationally inattentive utility maximization by a large margin, as displayed in Table I. Our findings are based on approximately 500 experiments on 5 widely used neural network architectures for image classification. This result establishes that the rationally inattentive utility maximization widely used to explain human decisions explains deep image classification remarkably well. This result is schematically shown in Fig. 2.
2. To aid visualization of our interpretable model, we provide a sparsity-enhanced decision test that computes the sparsest utility function and information acquisition cost which rationalizes deep CNN decisions. The sparsest solution yields a parsimonious representation of hundreds of thousands of layer weights of the deep CNNs in terms of a few hundred parameters. The utility function of the sparsest interpretable model also induces a useful preference ordering amongst the set of hypotheses (image labels) considered by the CNN; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated. The sparsity results for various deep CNN architectures are displayed in Table II and Fig. 3.
3. Our final result demonstrates the usefulness of our interpretable model. We show that, via interpolation, the

interpretable model computed from CNN decisions can predict the classification accuracy of a CNN trained with arbitrary parameters with high accuracy. The prediction results are displayed in Table III. Finally, we propose Algorithm 3 that uses the reconstructed utility functions and costs to *predict* the label of an image (and mimicing a constrained Bayesian utility maximizer in action) hence providing a complete economics-based interpretable model for deep image classification.

The above results are backed by approximately 500 experiments performed on several deep CNN architectures on the CIFAR-10 [37] image dataset over 3 learning rates, 200 training epochs, and 20 values of noise variance for corrupting the original images. The first two results use deep CNN decisions aggregated over varying training epochs. The third (prediction) result uses deep CNN decisions trained on noisy image datasets parameterized by the noise variance.[3] Also, Appendix D contains numerical experiments for interpretable deep image classification using Vision Transformer (ViT) architecture on large image datasets, namely, Tiny-Imagenet and CIFAR-100. We conduct our experiments for 80 image classes and 5 trained neural networks, where the neural networks differ only with respect to 1 training parameter. Experiment 1 varies the variance of the noise added to the training images, and experiment 2 varies the training epochs.[4]

II. Bayesian Revealed preference with Rational Inattention

This section describes the key ideas behind Bayesian revealed preference. Despite the abstract formulation below, the reader should keep in mind the deep learning context. In Sec. III, we will use Bayesian revealed preference theory to construct an interpretable deep learning representation by showing that deep CNNs are equivalent to rationally inattentive Bayesian utility maximizers.

A. Utility Maximization with Rational Inattention (UMRI)

Bayesian revealed preference aims to determine if the decisions of a Bayesian agent are consistent with expected utility maximization subject to a rational inattention sensing cost. We start by describing the **u**tility **m**aximization model with **r**ational **i**nattention (henceforth called UMRI) for a ***collection*** of Bayesian decision makers/agents.

Abstractly, the UMRI model is parameterized by the tuple

$$\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, C, \{\alpha_k, u_k, k \in \mathcal{K}\}). \tag{1}$$

With respect to the abstract parametrization of the UMRI model for a collection of Bayesian agents, the following elements constitute the tuple Θ defined in (1).

Agents: $\mathcal{K} = \{1, 2, \ldots, K\}$ $(K \geq 2)$ indexes the finite set of Bayesian agents.

State: $\mathcal{X}$ is the finite set of ground truths with prior probability distribution π_0. With respect to our image classification context, $\mathcal{X} = \{1, 2, \ldots 10\}$ is the set of image classes in the CIFAR-10 dataset and π_0 is the empirical probability distribution of the image classes in the test dataset of CIFAR-10.

Observation and attention strategy: Agent $k \in \mathcal{K}$ chooses attention strategy $\alpha_k : \mathcal{X} \to \Delta(\mathcal{Y})$, a stochastic mapping from $\mathcal{X}$ to a finite set of observations $\mathcal{Y}$. Given state x and attention strategy α_k, the agent samples observation y with probability $\alpha_k(y|x)$. The agent then computes the posterior probability distribution $p(x|y)$ via Bayes formula as

$$p(x|y) = \frac{\pi_0(x)\alpha_k(y|x)}{\sum_{x' \in \mathcal{X}} \pi_0(x')\alpha_k(y|x')}. \tag{2}$$

The observation and attention strategy are latent variables that abstractly represent the learned feature representations in the deep image classification context. Bayesian revealed preference theory tests their existence via the convex feasibility test in Theorem 1 below.

Action: Agent $k \in \mathcal{K}$ chooses action a from a finite set of actions $\mathcal{A}$ after computing the posterior probability distribution $p(x|y)$. In the image classification context, a is the image class predicted by the neural network, hence $\mathcal{A} = \mathcal{X}$.

Utility function: Agent $k \in \mathcal{K}$ has a utility function[5] $u_k(x, a) \in \mathbb{R}^+$, $x \in \mathcal{X}, a \in \mathcal{A}$ and aims to maximize its expected value, with the expectation taken wrt the random state x and random observation y. A key feature in our approach is to show that the utility function rationalizes the decisions of the deep CNNs (made precise in Definition 1).

Information Acquisition Cost: The information acquisition cost $C(\alpha, \pi_0) \in \mathbb{R}^+$ depends on attention strategy α and prior pmf π_0. It is the sensing cost the agent incurs to estimate the underlying state (2). In the context of machine learning, $C(\cdot)$ abstractly captures the 'learning' cost incurred during the training of the deep neural networks. In rational inattention theory from behavioral economics, a higher information acquisition cost is incurred for more accurate attention strategies (equivalently, more accurate state estimates (2) given observation y). We refer the reader to the influential work of [1], [2].

Each Bayesian agent $k \in \mathcal{K}$, aims to maximize its expected utility while minimizing its cost of information acquisition. Hence, the action a given observation y, and attention strategy α_k are chosen as follows:

Definition 1 (Rationally Inattentive Utility Maximization)**.** *Consider a collection of Bayesian agents $\mathcal{K}$ parameterized by Θ in (1) under the UMRI model. Then,*

(a) **Expected Utility Maximization:** *Given posterior probability distribution $p(x|y)$, every agent $k \in \mathcal{K}$ chooses action a that maximizes its expected utility. That is, with $\mathbb{E}$ denoting mathematical expectation, the action a satisfies:*

$$a \in \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \ \mathbb{E}_x\{u_k(x, a')|y\} = \sum_{x \in \mathcal{X}} p(x|y)u_k(x, a'). \tag{3}$$

(b) **Attention Strategy Rationality:** *For agent k, the attention strategy α_k optimally trades off between maximizing the*

[3]The neural network classification accuracy as learning rates and training epochs are varied can be downloaded from `zerenzhang2022.github.io`

[4]All numerical results are completely reproducible and can be accessed from our public repo `github.com/aditj/extrasimulationsdlri`.

[5]Strictly speaking, $u_k \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{A}|}$ is a matrix with non-negative entries, with $u_k(x, a)$ denoting the entry in the x^{th} row and a^{th} column. The term 'utility function' is prevalent in micro-economics literature and refers to a matrix. In this paper, we borrow the micro-economics terminology and refer to the matrix u_k as the utility function.

expected utility and minimizing the information acquisition cost.

$$\alpha_k \in \underset{\alpha'}{\operatorname{argmax}}\, \mathbb{E}_y\{\max_{a\in\mathcal{A}} \mathbb{E}_x\{u_k(x,a)|y\}\} - C(\alpha', \pi_0). \quad (4)$$

Eq. 3,4 in Definition 1 constitute a nested optimization problem. The lower-level optimization task is to choose the the 'best' action for any observation y based on the computed posterior belief of the state. The upper-level optimization task is to sample the observations optimally by choosing the 'best' attention strategy.

Remark. The multiple Bayesian agents in Θ have the same state space $\mathcal{X}$, observation space $\mathcal{Y}$, action space $\mathcal{A}$, prior π_0 and cost of information acquisition C, but only differ in their utility functions. Bayesian revealed preference theory relies on this crucial constraint on the optimization variables in (3), (4) for detecting optimal behavior in a finite number of agents.

B. Bayesian Revealed Preference (BRP) Test for Rationally Inattentive Utility Maximization

Having described the UMRI model (collection of rationally inattentive utility maximizers), we are now ready to state our key result. Theorem 1 below says that the decisions of a collection of Bayesian agents is rationalized by a UMRI tuple Θ *if and only if* a set of convex inequalities have a feasible solution. These inequalities comprise our **B**ayesian **R**evealed **P**reference (henceforth called BRP) test for rationally inattentive utility maximization.

For notational convenience, the decisions of the Bayesian agents in the UMRI model are compacted into the dataset $\mathbb{D}$ defined as:

$$\mathbb{D} = \{\pi_0, p_k(a|x), x \in \mathcal{X}, a \in \mathcal{A}, k \in \mathcal{K}\}. \quad (5)$$

In (5), $\pi_0 \in \Delta^{|\mathcal{X}|-1}$ denotes the prior pmf over the set of states $\mathcal{X}$ in Θ (1). The variable $p_k(a|x)$ is the conditional probability that agent $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ takes action a given state x. $\mathbb{D}$ characterizes the input-output behavior of the collection of Bayesian agents and serves as the input for BRP feasibility test described below.

Theorem 1 (BRP Test for Rationally Inattentive Utility Maximization [7])**.** *Given the dataset $\mathbb{D}$ (5) obtained from a collection of Bayesian agents $\mathcal{K}$. Then,*

1. *Existence: There exists a UMRI tuple $\Theta(\mathbb{D})$ (1) that rationalizes dataset $\mathbb{D}$ if and only if there exists a feasible solution that satisfies the set of convex inequalities*

$$BRP(\mathbb{D}, \{u_k, c_k\}_{k=1}^K) \leq \mathbf{0},\ u_k \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{A}|},\ c_k > 0. \quad (6)$$

In (6), $BRP(\cdot)$ corresponds to a set of convex (in the variables $\{u_k, c_k\}_{k=1}^K$) inequalities, stated in Algorithm 1.

2. *Reconstruction: Given a feasible solution $\{u_k, c_k\}_{k=1}^K$ to $BRP(\mathbb{D}, \cdot)$, u_k is the k^{th} Bayesian agent's utility function in the feasible model tuple $\Theta(\mathbb{D})$. The feasible cost of information acquisition C in $\Theta(\mathbb{D})$ is defined in terms of c_k as:*

$$C(\alpha) = \max_{k\in\mathcal{K}}\ c_k\ + \sum_a \max_{b\in\mathcal{A}} \sum_x p(x,a)u_k(x,b) - \sum_{x,a} p_k(x,a)u_k(x,a), \quad (7)$$

where $\alpha = \{p(a|x), a \in \mathcal{A}, x \in \mathcal{X}\} \in \Delta(\mathcal{A})^{|\mathcal{X}|}$.

The proof of Theorem 1 is in Appendix A. Before launching into a detailed discussion, we stress the "iff" in Theorem 1. Put simply: if the inequalities in (6) are not feasible, then the Bayesian agents that generate the dataset $\mathbb{D}$ are not rationally inattentive utility maximizers. If (6) has a feasible solution[6], then there exists a reconstructable family of viable utility functions and information acquisition costs that rationalize $\mathbb{D}$[7]. A key feature of Theorem 1 is that the estimated utilities (and information costs) are set-valued; every utility and cost function in the feasible set explains $\mathbb{D}$ equally well. The estimated UMRI model parameters are set-valued due to the finite number of Bayesian agents whose decisions constitute the dataset $\mathbb{D}$. The estimated parameter set converges to a point if and only if the inequality (6) holds as $|\mathcal{K}| \to \infty$.

Computational Aspects of BRP Test. Suppose the dataset $\mathbb{D}$ is obtained from K Bayesian agents. Then, BRP($\mathbb{D}$) comprises a feasibility test with $K\ (|\mathcal{X}||\mathcal{A}| + 1)$ free variables and $K^2 + K\ (|\mathcal{A}|^2 - |\mathcal{A}| - 1)$ convex inequalities. Thus, the number of free variables and inequalities in the BRP feasibility test scale linearly and quadratically, respectively, with the number of observed Bayesian agents.

Single Utility BRP (S-BRP). In Algorithm 4 in the appendix, we define a second set of inequalities S-BRP. The only difference between BRP and S-BRP is the number of variables. While BRP reconstructs a set of distinct utility functions indexed by the agent that rationalizes dataset $\mathbb{D}$, S-BRP assumes a single utility function but distinct Lagrange multipliers for the expected utility for all agents. Hence, S-BRP can be viewed as a more restrictive version of BRP.

Algorithm 1 BRP Convex Feasibility Test of Theorem 1

Require: Dataset $\mathbb{D} = \{\pi_0, p_k(a|x), x \in \mathcal{X}, a \in \mathcal{A}, k \in \mathcal{K}\}$ from a collection of Bayesian agents $\mathcal{K}$.

Find: Positive reals $c_k,\ u_k(x,a) \in (0,1]$ for all $x \in \mathcal{X}$, $a \in \mathcal{A},\ k \in \mathcal{K}$ that satisfy the following inequalities:

$$\underline{\textbf{NIAS}}: \sum_x p_k(x|a)\ (u_k(x,b) - u_k(x,a)) \leq 0, \quad \forall a, b \in \mathcal{A},\ k \in \mathcal{K}, \quad (8)$$

$$\underline{\textbf{NIAC}}: \sum_a \left(\max_b \sum_x p_j(x,a)u_k(x,b)\right) - c_j - \sum_{x,a} p_k(x,a)u_k(x,a) + c_k \leq 0,\ \forall j, k \in \mathcal{K}, \quad (9)$$

where $p_k(x,a) = \pi_0(x)p_k(a|x),\ p_k(x|a) = \frac{p_k(x,a)}{\sum_{x'} p_k(x',a)}$.

Return: Set of feasible utility functions u_k and information acquisition costs c_k incurred by agents $k \in \mathcal{K}$.

[6]Although expressed differently, the NIAC condition (9) in Algorithm 1 is equivalent to that in the original work of [7, Theorem 1]. In [7], the NIAC condition does not involve the information cost terms as feasible variables.

[7]In terms of interpretable deep learning, of all parameters in the UMRI tuple, we are only interested in the utility functions of the agents and the cost of information acquisition, since the remaining parameters can be inferred from the decision dataset $\mathbb{D}$.

C. BRP test and Interpretable Deep Image Classification

We now discuss how the above BRP test relates to interpretable image classification using deep CNNs. The BRP convex feasibility test in Theorem 1 comprises two sets of inequalities, namely, the *NIAS* (No-Improving-Action-Switches) (8) and *NIAC* (No-Improving-Action-Cycles) (9) inequalities (Algorithm 1). NIAS ensures that the agent takes the best action given a posterior pmf. NIAC ensures that every agent chooses the best attention strategy. BRP test checks if there exist K utility functions and K positive reals that, together with $\mathbb{D}$, satisfy the NIAS and NIAC inequalities.

Toy Example with 2 CNNs: The following discussion gives additional insight into our approach. Consider the simplest case involving two trained deep CNNs N_1 and N_2; so $\mathcal{K} = \{1, 2\}$ in the above notation. Assume N_1 and N_2 have the same network architecture. Suppose an analyst observes that N_1 makes accurate decisions on a rich input image dataset while N_2 makes less accurate decisions on the same dataset.

Our UMRI model first abstracts the accuracy of the feature representations of the input image data learned by N_1 and N_2 via attention strategies α_1 and α_2 in (4). Second, the information acquisition cost function $C(\cdot)$ abstracts the computational resources expended for learning the representations. The rationale is that learning an accurate latent feature representation is costly, and this is abstracted by the information acquisition cost.

Let the training cost incurred by N_1 and N_2 be $C(\alpha_1)$ and $C(\alpha_2)$ respectively. If the decisions of N_1 and N_2 can be explained by the UMRI model (and Theorem 1 above will give necessary and sufficient conditions for this), then there exist utility functions u_1 and u_2 for N_1 and N_2, that satisfy:

$$\mathbb{E}_{\alpha_i}\{u_i\} - C(\alpha_i) \geq \mathbb{E}_{\alpha_j}\{u_i\} - C(\alpha_j), \;\; i, j \in \{1, 2\} \quad (10)$$

The above inequality says that CNNs N_1 and N_2 would be worse off (in an expected utility sense) if they make decisions based on swapping each other's learned representations. That is, both N_1 and N_2 learn the 'best' feature representation of the input images given their training parameters.

Discussion: (i) *Parsimonious Interpretable Representation of deep CNNs.* In the deep image classification context, due to the UMRI model's parsimonious parametrization in (1), the decisions of K CNNs can be rationalized by just K utility functions and an information acquisition cost function, thus bypassing the need of several million parameters to describe the deep CNNs.

(ii) *Identifiability.* The BRP feasibility test requires the dataset $\mathbb{D}$ to be generated from $K > 2$ Bayesian agents. If $K = 1$, then (6) holds trivially since any information acquisition cost satisfies the convex inequalities of BRP. Another intuitive way of motivating a collection of agents for the BRP is as follows. Reconstructing a feasible UMRI model tuple Θ that rationalizes the decisions of the deep CNNs is analogous to fitting a line to a finite number of points. One can fit infinitely many lines through a single point. The task becomes non-trivial if the number of points exceeds 2. In the Bayesian revealed preference context, the points correspond to the decisions from each Bayesian agent. The slope and intercept of the fitted line, in our case, corresponds to the utility function and cost of information acquisition that rationalize the agent decisions.

(iii) *Relative Optimality implies Global Optimality.* In the setting involving $K > 2$ deep CNNs (agents), the NIAS and NIAC inequalities of BRP test check for relative optimality - *given utility function u_k, does deep CNN k performs at least as well as any other observed deep CNN in $\mathcal{K}\backslash\{k\}$?* Clearly, testing for relative optimality is weaker than testing for global optimality (4) which ideally requires access to decisions from an infinite number of deep CNNs. Setting the cost of information acquisition as a free variable bridges this gap. The proof of Theorem 1 shows that if the deep CNN decisions satisfy relative optimality, then there exists a cost of information acquisition such that the decisions are globally optimal. That is, Theorem 1 ensures relative optimality is sufficient for global optimality.

(iii) *Generalization of [7].* Theorem 1 generalizes [7, Theorem 1] in two ways. (1) In [7], the utilities u_k in UMRI model tuple Θ are assumed known, and only the information acquisition costs c_k are estimated, whereas Theorem 1 estimates both parameters. (2) The expression for the reconstructed model tuple $\Theta(\mathbb{D})$ is novel; the discussion in [7] is only confined to the existence of such a tuple.

(vi) *Single Utility UMRI (S-UMRI).* In Appendix B, we propose a sparse version of UMRI , namely, the S-UMRI model in (23). The key distinction of this model is that all agents have the same utility function u and thus can be represented with substantially fewer parameters. In complete analogy to Theorem 1, we outline a decision test in Theorem 3 that states necessary and sufficient conditions for agent decisions to be consistent with the S-UMRI model of rationally inattentive utility maximization. We discuss this sparse parametrization in the appendix so as not to interrupt the flow of the main text.

(vii) *Degenerate solution to BRP test.* The degenerate utility function of all zeros and cost of information acquisition $C = 0$ trivially satisfies the BRP tests and lie at the boundary of the feasible set of parameters.

Summary: This section formulated an economics-based decision-making model. Since this model may not be familiar to a machine learning reader, we summarize the main ideas. We introduced the rationally inattentive utility maximization model, namely, the UMRI model for a collection of Bayesian agents (decision makers). Our main result Theorem 1 outlines a decision test BRP for rationally inattentive utility maximization given decisions from a collection of agents. This BRP test comprises a set of convex inequalities that have a feasible solution *if and only if* the collection of agents are rationally inattentive utility maximizers. Theorem 1 also provides an explicit reconstruction of the feasible UMRI model parameters that rationalize input agent decisions. The set of feasible utility functions and information acquisition costs thus parsimoniously explain the decisions generated by the Bayesian agents. In Appendix B, we propose a single utility version of the UMRI model with fewer parameters. Due to fewer parameters, the decision test for this sparse model, given in Theorem 3, is computationally less expensive yet more restrictive than the

BRP test for rationality in Theorem 1.

The rest of the paper focuses on computing interpretable UMRI models that rationalize deep CNN decisions. We will investigate through extensive experiments how well the UMRI fits the deep CNN decisions via robustness tests. We will also investigate how well the computed interpretable models, namely, UMRI and S-UMRI , predict the deep CNNs' decisions when the training parameters are varied.

III. Bayesian Revealed Preference explains CIFAR-10 Image Classification by Deep CNNs

The experimental results in this section are divided into two parts: First, we show that the deep CNNs decisions pass the BRP and S-BRP tests formulated in Theorems 1 and 3 by a large margin. This implies that the rationally inattentive utility maximization model is a robust fit to the deep CNN decisions.

Our second result demonstrates an application of the reconstructed interpretable model. Training datasets are often noisy. We show that in such a noisy setting, the reconstructed interpretable model from Theorem 1 can accurately predict (with accuracy exceeding 94%) the image classification performance of the deep CNNs. This bypasses the need to train the deep CNN for various noise variances that corrupt the training dataset.

Experimental Setup: Deep CNN Architectures, Training Parameters and Construction of Dataset

Image Dataset. For our numerical experiments, we trained and validated the deep CNNs using the CIFAR-10 benchmark image dataset [37]. This public dataset consists of 60000 32x32 color images in 10 distinct classes (for example, airplane, automobile, ship, cat, dog, etc.), with 6000 images per class. There are 50000 training images and 10000 test images. We will use the terms image classes and image labels interchangeably.[8]

Network Architecture and Training Parameters. In this paper, we use 5 well-known deep CNN architectures for our experiments. 1. LeNet [38], 2. AlexNet [39] 3. VGG16 [40] 4. ResNet-50 [22] 5. Network-in-Network (NiN) [41] The deep CNNs are trained and validated on the CIFAR-10 image dataset, using 3 learning rate schedules, namely, L.R. 1, L.R. 2 and L.R. 3. All 3 schedules use the RMSprop optimizer [42] with the decay parameter and maximum training epochs (full passes of the training dataset) set to 10^{-6} and 200, respectively, and initial step size set to 0.01. The step size is halved every $20, 30, 40$ epochs, respectively, for L.R. 1, 2 and 3.

Relation to Bayesian revealed preference. We now relate the deep CNN setup to the Bayesian revealed preference framework in Sec. II. For each CNN architecture, we use the decisions of $K = 20$ CNNs, i.e., 20 Bayesian agents in the terminology of Sec. II, for our BRP and S-BRP decision tests. The CNN decisions from K CNNs on the test image dataset of CIFAR-10 are aggregated into dataset $\mathbb{D}$ (5). The results of the decision tests are discussed below. In the deep image classification context, the parameter $p_k(a|x)$ in (5) is the probability that the k^{th} deep CNN classifies an image from category x into category a in the CIFAR-10 test image dataset. The prior π_0 in $\mathbb{D}$ (5) is the empirical pmf over the set of image categories in the CIFAR-10 test dataset. Constructing $\mathbb{D}$ from raw CNN decisions is discussed in Appendix C.

[8]Our experiments are confined to the CIFAR-10 dataset in the main text for clarity of exposition. Our approach to interpretable deep learning can be easily extended to richer benchmark image datasets like ImageNet and CIFAR-100 (that comprise over 100 image labels); see Appendix D for numerical results from additional experiments on larger image datasets.

A. BRP and S-BRP Tests for deep CNN datasets. Results and Insights

Network Architecture	Learning Rate	$\mathcal{R}_{\text{BRP}}$ $(\times 10^{-4})$	$\mathcal{R}_{\text{S-BRP}}$ $(\times 10^{-4})$
LeNet	L. R. 1	30.34	4.72
	L. R. 2	35.14	4.65
	L. R. 3	37.97	5.11
AlexNet	L. R. 1	32.10	3.21
	L. R. 2	34.98	3.91
	L. R. 3	40.60	4.62
VGG16	L. R. 1	96.36	4.09
	L. R. 2	107.4	4.02
	L. R. 3	119.8	4.44
ResNet-50	L. R. 1	126.2	2.82
	L. R. 2	129.2	3.45
	L. R. 3	132.3	3.83
Network-In-Network (NiN)	L. R. 1	108.3	3.59
	L. R. 2	132.1	3.36
	L. R. 3	149.1	5.57

TABLE I: How does increasing the number of degrees of freedom of the interpretable model improve robustness of fit to the CNN decisions? We see that $\mathcal{R}_{\text{BRP}}$ (11) is substantially higher (by an order of magnitude) than $\mathcal{R}_{\text{S-BRP}}$ (12) for all CNN architectures. We conclude that the UMRI model fits CNN decisions substantially better than the S-UMRI model, but with larger computing cost for evaluating the parameters of the interpretable model. Thus, if there are no computational constraints, we recommend using the UMRI model for interpreting CNN decisions.

A. Robustness Results on Deep CNN datasets: Our first key result is that image classifications of all 5 deep CNN architectures listed in Sec. III pass the BRP and S-BRP tests by a large margin. The results are tabulated in Table I. The robustness values $\mathcal{R}_{\text{BRP}}$ and $\mathcal{R}_{\text{S-BRP}}$ in Table I are defined in Definition 2 below which formalizes the notion of margin for the decision tests.

Definition 2 (Robustness (Goodness-of-fit) of BRP and S-BRP Tests.)**.** *Given dataset $\mathbb{D}$ (5) aggregated from a collection of Bayesian agents, $\mathcal{R}_{BRP}(\mathbb{D})$ and $\mathcal{R}_{S\text{-}BRP}(\mathbb{D})$ measure the largest perturbation so that $\mathbb{D}$ passes the BRP and S-BRP decision*

Network Architecture	Learning Rate (L.R.)	airplane	auto	bird	cat	deer	dog	frog	horse	ship	truck
LeNet	L.R. 1	17.61	3.55	20.06	1.88	17.19	21.42	42.00	27.79	1.91	9.55
	L.R. 2	4.13	5.20	7.82	1.90	13.18	18.66	23.84	8.16	2.48	2.47
	L.R. 3	10.79	8.27	18.62	22.67	19.91	25.01	47.71	73.52	2.65	1.01
AlexNet	L.R. 1	210.78	41.84	49.77	59.71	51.24	68.31	83.94	211.61	60.43	125.73
	L.R. 2	85.51	47.89	17.38	1.00	25.34	202.78	21.30	35.01	533.62	248.57
	L.R. 3	18.00	49.55	58.25	28.31	135.54	29.24	224.91	214.51	8.29	264.20
VGG16	L.R. 1	164.48	154.77	15.42	33.67	6.28	123.89	62.83	26.21	1.43	170.69
	L.R. 2	88.73	154.10	45.63	297.61	131.08	136.52	57.34	229.80	145.99	11.90
	L.R. 3	24.33	10.78	93.90	11.11	91.96	56.64	77.30	110.60	20.28	17.09
ResNet-50	L.R. 1	50.83	17.55	16.09	4.66	17.92	3.67	4.92	3.95	15.46	4.88
	L.R. 2	7.51	8.40	72.70	30.72	32.43	83.65	221.27	74.59	99.04	20.51
	L.R. 3	14.61	367.59	31.61	9.20	16.35	11.58	41.44	243.95	222.67	483.91
Network-in-Network	L.R. 1	5.02	30.95	9.91	71.38	63.69	45.88	31.39	67.86	17.03	21.41
	L.R. 2	40.17	60.32	4.40	55.67	95.02	88.72	91.15	15.98	176.75	10.27
	L.R. 3	10.47	75.32	55.97	24.17	17.41	8.94	23.02	71.27	29.94	80.91

TABLE II: The utility function of the sparsest interpretable model is a diagonal matrix. The diagonal elements yield a natural preference ordering amongst the set of image classes (classification hypotheses). For example, consider the VGG16 architecture trained using learning rate 1 (third row, first sub-row of table). The maximum utility is for trucks (170.69, last column) and the minimum is for ships (1.43, second last column). This shows the sparsest interpretable model induces the following preference ordering for the VGG16 architecture: classifying trucks correctly is prioritized 100 times more than classifying ships. Such a preference ordering is not explicitly generated by a CNN.

tests:

$$\mathcal{R}_{BRP}(\mathbb{D}) = \max_{\varepsilon>0} \frac{\varepsilon\ K}{\sum_{k=1}^{K} \|u_k\|_2^2},\ BRP(\mathbb{D}, \{u_k, c_k\}_{k=1}^{K}) \leq -\varepsilon. \tag{11}$$

$$\mathcal{R}_{S\text{-}BRP}(\mathbb{D}) = \max_{\varepsilon>0} \frac{\varepsilon}{\|u\|_2^2}, S\text{-}BRP(\mathbb{D}, u, \{c_k, \lambda_k\}_{k=1}^{K}) \leq -\varepsilon. \tag{12}$$

In Definition 2, robustness values $\mathcal{R}_{\text{BRP}}$ and $\mathcal{R}_{\text{S-BRP}}$ measure, respectively, the smallest perturbation needed for $\mathbb{D}$ to fail the BRP and S-BRP decisions tests. Put differently, the variable ε measures how well the BRP and S-BRP inequalities are satisfied given dataset $\mathbb{D}$. The higher the value of feasible ϵ that satisfies the constraint in the optimization problems (11) and (12), the better the corresponding utilities and costs explain the dataset $\mathbb{D}$, hence indicating a better fit of the UMRI and S-UMRI model to neural networks' performance. This 'max-margin' philosophy is prevalent in both IRL [43] and revealed preference [44] literature. Both $\mathcal{R}_{\text{BRP}}$ and $\mathcal{R}_{\text{S-BRP}}$ are normalized wrt the row-wise $\mathcal{L}_2$ norm of the feasible utility functions. Higher robustness values imply a better fit of the UMRI, S-UMRI models to the decision dataset [9].

Discussion and Insights. Robustness Results of Table I:
(i) *Deep CNN dataset:* The deep CNN datasets used for the robustness tests (11), (12) comprise decisions of $K = 20$ deep CNNs for every network architecture, where CNN k was trained for 10 k training epochs, $k = 1, 2, \ldots, K$.
(ii) *Comparison between $\mathcal{R}_{BRP}$ and $\mathcal{R}_{S\text{-}BRP}$ values for deep CNN datasets:* The average value of $\mathcal{R}_{\text{S-BRP}}$ (12) over all 3 learning rate schedules and 5 network architectures was found to be 4.09×10^{-4}. In contrast, the average value of $\mathcal{R}_{\text{BRP}}$ (11) was found to be 87.45×10^{-4}, almost 20 times the average value of $\mathcal{R}_{\text{S-BRP}}$. This result shows that the UMRI model fits deep CNN decisions substantially better than the S-UMRI model. This result is expected since S-UMRI is parameterized using much fewer variables compared to the UMRI and hence, S-BRP test is more restrictive than BRP.
(iii) *Sensitivity of $\mathcal{R}_{BRP}, \mathcal{R}_{S\text{-}BRP}$ to Network Architecture:* The average value of $\mathcal{R}_{\text{BRP}}$ is 122.29×10^{-4} for the LeNet and AlexNet architectures, which is approximately 3.5 times the the average value of $\mathcal{R}_{\text{BRP}}$ for the VGG16, ResNet-50 and NiN architectures which is 35.18×10^{-4}. The variation of $\mathcal{R}_{\text{S-BRP}}$ with network architecture is negligible compared to $\mathcal{R}_{\text{S-BRP}}$. This shows the robustness test for UMRI model is more sensitive to network architecture compared to that for the S-UMRI model.
(iv) *Computational aspects of $\mathcal{R}_{BRP}$ and $\mathcal{R}_{S\text{-}BRP}$.* The computation time for $\mathcal{R}_{\text{BRP}}$ is almost 30 times that for $\mathcal{R}_{\text{S-BRP}}$. This is expected since the UMRI model is parameterized by K utility functions compared to a single utility function in S-UMRI.

B. Sparsity-enhanced Interpretable Model: Our next task is to determine the sparsest possible interpretable model that satisfies the decision tests BRP and S-BRP. The motivation is three fold:

1) The sparsest interpretable model explains the deep CNN decisions using the fewest number of parameters.
2) The sparsest interpretable model induces a useful preference ordering amongst the set of hypotheses (image labels) considered by the CNN; for example, how much

[9]The robustness value for the non-informative dataset of uniformly distributed pmfs is 0. Hence, the robustness value measures the informativeness of the attention strategies in $\mathbb{D}$ relative to the uniform probability distribution.

additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated.

3) Third, the sparsest solution is a point valued estimate. Recall the BRP and S-BRP decision tests yield a set-valued estimate of feasible utility functions and cost of information acquisition that explain the deep CNN datasets. While every element in the set explains the dataset equally well, it is useful to have a single representative point.

Theorem 2 below computes the sparsest utility function out of all feasible utility functions.

Theorem 2 (Sparsity Enhanced BRP and S-BRP Tests for Deep CNN datasets)**.** *Given dataset* $\mathbb{D}$ (5) *from a collection of* K *Bayesian agents. The sparsest solutions to the BRP and S-BRP tests minimize the sum of row-wise* $\mathcal{L}_1$ *norm of the feasible utility functions of the* K *agents that generate* $\mathbb{D}$*.*

$$(u_{1:K})^* = \underset{u_{1:K}}{\operatorname{argmin}} \sum_{k=1}^{K} \|u_k\|_1, \textit{BRP}(\mathbb{D}, \cdot) \leq \mathbf{0}, \ \sum_{k=1}^{K} \|u_k\|_2^2 = K.$$
$$u^* = \underset{u}{\operatorname{argmin}} \|u\|_1, \textit{S-BRP}(\mathbb{D}, \cdot) \leq \mathbf{0}, \ \|u\|_2^2 = 1. \tag{13}$$

where $\|\cdot\|_1$ *denotes the row-wise* $\mathcal{L}_1$ *norm.*

Results and Discussion. Sparsity Test for deep CNN datasets: The sparsest utility function from the S-BRP test are tabulated in Table II for all 5 deep CNN architectures. The corresponding information acquisition cost for all 5 architectures averaged over learning rates $1, 2, 3$ are shown in Fig. 3. Together, the sparsest utility and information cost constitute the sparsest S-UMRI interpretable model[10] for the deep CNN decisions.

(i) *Preference ordering induced from sparsest utility.* The sparsest utility function for the S-UMRI model induces a useful preference ordering among the predicted image classes. That is, they measure how the deep CNN's priority for accurate classification varies across image classes. For instance, consider the VGG16 architecture trained using learning rate schedule 1. Of all image categories, the maximum utility is observed for trucks (170.69) and the minimum for ships (1.43). This shows the VGG16 architecture prioritizes classifying trucks correctly about 100 times more than classifying ships.

(ii) *Penalty for learning image features accurately.* The computed information acquisition costs in Fig. 3 can be understood as the training cost the CNN incurs to learn latent image features accurately. The interpretable model cannot explain the variation in CNN classification accuracy versus variation in training parameters without an information acquisition cost. From Fig. 3, we can conclude that learning accurate image features is the most and least costly, respectively, for the AlexNet and ResNet architectures, respectively.

[10]For brevity, we have only included the sparsity results for the S-UMRI model. The sparsest utility functions of the UMRI model that explains deep CNN decisions are included in our public GitHub repository that contains all test results and codes.

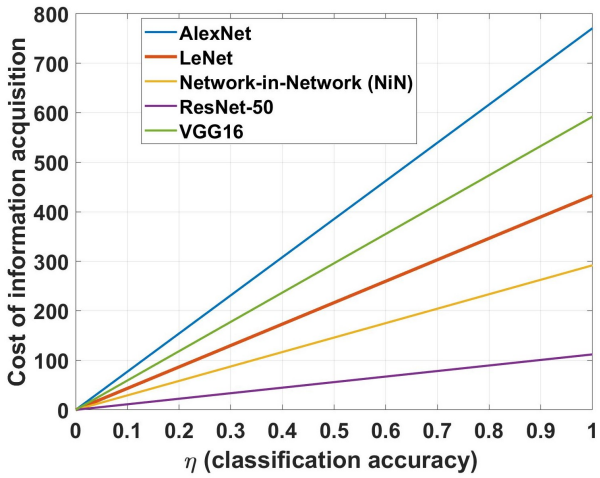


Fig. 3: The figure illustrates an important property of our approach to interpretable deep learning: in addition to the utility function (Table II), we also need a rational inattention term (cost of learning latent image features) to explain CNN decisions. Put differently, we cannot explain the variation in CNN classification accuracy versus variation in training parameters without an information acquisition cost. The figure displays the information acquisition cost C (7) evaluated for the sparsest interpretable model. We also observe that learning accurate image features is most expensive for AlexNet, and least expensive for ResNet architectures.

B. Predicting deep CNN classification accuracy using our Interpretable Models

Training datasets are often noisy; for example, [45] considers noisy datasets for hand-written character recognition. We now exploit the proposed interpretable model to predict how the deep CNN will perform with a noisy training dataset without actually implementing the deep CNN.

Our predictive procedure is as follows. We first train the CNNs on noisy datasets that are generated by adding simulated Gaussian noise with noise variances chosen from a finite set. [11] Then given the CNN decisions, we compute our interpretable model over this finite set of noise variances. Finally, to predict how the CNN will perform for a noise variance not in the set, we *interpolate* the utility function of the interpretable model at this noise variance. Then given the interpolated utility function and information acquisition cost from our interpretable model, the predicted classification performance is computed by solving convex optimization problem (4). The above procedure is formalized in Algorithm 2. Hence, our interpretable model serves as a computationally efficient method for predicting the performance of a CNN without implementing the CNN. The interpretable model can be viewed as a low-dimension projection of the high-dimension CNN with predictive accuracy exceeding 94%.

Remark. An alternative procedure is to directly interpolate the performance over the space of CNN weights (several hundreds of thousands). Due to the high dimensionality, this is

[11]Injecting artificial noise in training datasets is also used in variational auto-encoders for robust feature learning [46], [47].

Network Architecture	airplane	auto	bird	cat	deer	dog	frog	horse	ship	truck
LeNet	0.042	0.042	0.041	0.027	0.046	0.025	0.049	0.034	0.040	0.042
AlexNet	0.025	0.031	0.034	0.021	0.046	0.032	0.049	0.039	0.045	0.036
VGG16	0.033	0.035	0.043	0.041	0.048	0.048	0.035	0.046	0.037	0.048
ResNet-50	0.030	0.031	0.027	0.031	0.020	0.027	0.040	0.015	0.023	0.024
Network-in-Network	0.051	0.029	0.025	0.028	0.056	0.059	0.030	0.058	0.045	0.036

TABLE III: How well does our interpretable model predict CNN classification accuracy? The table displays the prediction error $\delta_\eta(x)$ defined in (14). Recall $\delta_\eta(x)$ is the error between the true CNN performance and the predicted performance using the interpretable model with Algorithm 2. The maximum error across all image classes and architectures was found to be 5.9%. Hence, our interpretable model predicts CNN classification performance with accuracy exceeding 94%.

an intractable interpolation. In comparison, interpolation over the utility functions in our interpretable model is over a few hundred variables.

Prediction Results of Algorithm 2 on Deep CNN Performance: Table III displays the prediction errors (difference between the true and predicted classification accuracy) for the deep CNNs for all 5 architectures and all image classes in CIFAR-10. For a fixed CNN architecture and noise variance $\eta > 0$, the prediction error $\delta_\eta(x)$ for image class x is defined as:

$$\delta_\eta(x) = |\hat{p}(x|x) - p_{\text{CNN}}(x|x)|. \tag{14}$$

In (14), $\hat{p}(\cdot|\cdot)$ is the predicted CNN performance generated from Algorithm 2 and $p_{\text{CNN}}(\cdot|\cdot)$ is the true CNN performance obtained by implementing the CNN. Recall that $p(x|x)$ is the probability that the CNN correctly classifies an image belonging to class x.

Algorithm 2 Predicting Deep CNN Classification Accuracy via the S-UMRI model using Theorem 2.

Require: Dataset $\mathbb{D}$ (28) from K deep CNNs from a fixed network architecture. The k^{th} CNN is trained on a noisy dataset intentionally perturbed by additive zero mean Gaussian noise on each pixel, with noise variance set to $\eta_k = 1 + 0.1 \times (k-1)$.

Step 1: *Constructing Interpretable Model.* The most robust utility functions $\{u_k^*\}_{k=1}^K$ and information acquisition cost C^* are computed by solving the following convex optimization problem.

$$\{u_k^*, c_k^*\}_{k=1}^K = \underset{u_{1:K}}{\text{argmax}} \frac{\varepsilon\ K}{\sum_{k=1}^K \|u_k\|_2^2},\ \text{BRP}(\mathbb{D}, \cdot) \leq -\varepsilon.$$
$$C^*(p(a|x)) = \max_{k=1} c_k^* + \sum_{x,a} \pi_0(x)(p(a|x) - p_k(a|x))u^*(x,a). \tag{15}$$

Step 2: *Predicting Classification Accuracy.* For an arbitrary noise variance $\eta \in [\eta_1, \eta_K]$, obtain index $g \in \mathbb{Z}_+,\ g \leq K$ such that $\eta \in [\eta_g, \eta_{g+1}]$. Then, the predicted classification accuracy $\hat{p}(a|x)$ for noise variance η is computed as follows:

$$\hat{p}(a|x) = \underset{p(a|x)}{\text{argmax}} \sum_a \max_b \sum_x \pi_0(x) p(a|x) \hat{u}(x,a) - C^*(p),$$
$$\hat{u} = 10 \times \{(\eta_{g+1} - \eta)u_g^* + (\eta - \eta_g)u_{g+1}^*\}. \tag{16}$$

Return: Predicted performance $\hat{p}(a|x)$ for noise variance η.

Discussion and Insights: (i) Our interpretable model can predict CNN classification performance at the image label level with high accuracy (see below).
(ii) The interpretable model (utility functions and information acquisition cost) for our predictive procedure (Algorithm 2) is evaluated on the set of noise variances $G_1 = \{1 + 0.1 \times (k-1), k = 1, 2, \ldots, 11\}$. The predictive procedure of Algorithm 2 is applied on the set of noise variances given by $G_2 = \{1.05 + 0.1 \times (k-1), k = 1, 2, \ldots, 10\}$. Table III displays the prediction errors $\delta_\eta(x)$ averaged over all $\eta \in G_2$.
(iii) From Table III, the prediction error $\delta_\eta(x)$ averaged over all image classes x for the 5 CNN architectures are:

1) LeNet - 0.038
2) AlexNet - 0.036
3) VGG16 - 0.041
4) ResNet-50 - 0.027
5) NiN- 0.035

So the least accuracy is 95.9%, and highest accuracy is 97.3%.
(iv) The prediction error averaged over the network architectures was observed to be minimum for image class 'cat' (98.1%) and maximum for image class 'deer' (95.7%) overall image classes.
(iv) *Statistical Similarity between Deep CNNs and Interpretable Model.* We computed the Kullback-Leibler (KL) divergence between the true and predicted classification performances $p_{\text{imp}}(a|x)$ and $\hat{p}(a|x)$. Recall $\hat{p}(a|x)$ is computed from the interpretable model via Algorithm 2 and $p_{\text{CNN}}(a|x)$ is obtained from the CNN. The KL divergence values for the 5 CNN architectures are:

1) LeNet - 0.015
2) AlexNet - 0.012
3) VGG16 - 0.016
4) Resnet-50 - 0.006
5) NiN - 0.018.

Thus, the decisions made by the deep CNNs are statistically similar to decisions generated by our interpretable model.

Remarks.

1) Although our numerical experiments only consider the CIFAR-10 image dataset, our results are straightforward to extend to larger and more granular datasets like CIFAR-100 [37] and ImageNet [48] at the cost of greater computational resources. In Appendix D, we describe additional experiments performed using the state-of-the-art vision transformer (ViT) [49] on larger image datasets like CIFAR-100 and Tiny-ImageNet. We construct two sets

of utility functions and associated information acquisition costs corresponding to the UMRI model. We also compute the goodness-of-fit of the UMRI model to the neural network dataset $\mathbb{D}$ (5).

2) *Predicting neural network behavior. Performance vs. Classification.* The rational inattention-based interpretable model (UMRI) takes as input a prior distribution over image labels and the confusion matrix generated from the trained neural network. Then, it yields a set of feasible utility functions and costs that rationalize the model inputs. In other words, the UMRI model explains the neural network's performance; see Fig. 1 for an illustration. Sec. III-B shows how an interpolation-based scheme can be used to *predict* the performance, or equivalently, the confusion matrix for neural networks trained wrt unknown learning parameters.

However, it is straightforward to formulate a *classification scheme* using the utility functions and costs generated by the interpretability test. Simply put, one only needs to execute the lower block diagram in Fig. 1 using the parameters computed from our interpretability algorithm[12] This is an interesting area of future research; we provide a brief outline in Algorithm 3.[13]

IV. Conclusions and Extensions

This paper proposed a data-driven micro-economics based system identification approach for interpretable deep classification. The key results stem from Bayesian revealed preference. By embedding deep image classification in a constrained Bayesian utility maximization framework, interpretable deep image classification is equivalent to set-valued system identification of an argmax non-linearity (in signal processing terms). Based on approximately 500 experiments on 5 popular CNN architectures, we showed that deep CNNs can be explained remarkably well by Bayesian utility maximization constrained by an information cost.

Our main results were the following:

1. Using the theory of Bayesian revealed preference, Theorem 1 gave a necessary and sufficient condition for the actions of a collection of decision makers to be consistent with rationally inattentive Bayesian utility maximization. We showed that deep CNNs operating on the CIFAR-10 dataset satisfy these necessary and sufficient conditions.

2. Next, we studied the robustness margin by which the deep CNNs satisfy Theorem 1; we found that the margins were sufficiently large, implying the robustness of the results. Our robustness results are summarized in Table II.

3. In Theorem 2, we constructed the sparsest interpretable model from the feasible set generated using Theorem 1. The sparsest interpretable model explains deep CNN decisions using the least number of parameters. The sparsest interpretable model introduces a useful preference ordering amongst the set of hypotheses (image labels) considered by the deep neural network; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated

4. Finally, we showed that our interpretable model can predict CNN performance with an accuracy exceeding 94%, and the decisions generated by our interpretable model are statistically similar to that of a deep CNN. At a more conceptual level, our results suggest that deep CNNs for image classification are equivalent to an economics-based constrained Bayesian decision system (used in micro-economics to model human decision-making).

5. We also conduct numerical experiments, namely, sparse utility estimation and robustness analysis, on larger image datasets using the state-of-the-art Vision Transform (ViT) neural network and display the reconstructed large dimensional utility values in the Appendix. To illustrate how the reconstructed interpretable model parameters, namely, utility function and costs, can be used to predict the label when provided an image, we also outline a pseudo-algorithm in Algorithm 3.

Extensions. An immediate extension of this work is to

Algorithm 3 UMRI-based Image Classification Protocol

Require: Image Dataset $\mathbb{D}_{\text{Image}} = \{\text{img}_i, x_i\}_{i=1}^{I}$ (x_i denotes the true image label of the i^{th} image), Trained neural networks indexed by k, $k \in \{1, 2, \ldots, K\}$.

Do:

(i) *Data Pre-processing.*

- Using any feature extraction method on the image dataset $\mathbb{D}_{\text{Image}}$, construct enriched dataset $\mathbb{D}_{\text{Image,Features}} = \{\text{img}_i, x_i, \tilde{x}_i\}_{i=1}^{I}$, where $\tilde{x}_i \in \tilde{\mathcal{X}}$ denotes the feature vector of image i.
- Compute feature priors $\tilde{\pi}_0$ and generate confusion matrices $\{p_k(a|\tilde{x})\}$ of the features, where $\tilde{x} \in \tilde{\mathcal{X}}$, $a \in \mathcal{A}$ and $\mathcal{A}$ denotes the set of predicted image labels.

(ii) *System Identification using Interpretability Test.* Compute the optimized robust point estimate of utility functions and sensing costs $\{u_k(\tilde{x}, a), C_k\}_{k=1}^{K}$ that rationalize the feature confusion matrices via (15).

(iii) *Function Evaluation of Rational Inattention Model.* Fix index i in the image dataset and trained neural network index k. Then:

- Sample action $a \sim p_k(\cdot|\tilde{x}_i)$, where $p_k(\cdot|\tilde{x})$ is the feature confusion matrix for feature $\tilde{x} \in \tilde{\mathcal{X}}$.
- Compute posterior belief $p_k(\cdot|a)$ corresponding to sampled action a using Bayes rule:

$$p_k(\tilde{x}|a) = \frac{\tilde{\pi}_0(\tilde{x})\ p_k(a|\tilde{x})}{\sum_{a' \in \mathcal{A}}\ \tilde{\pi}_0(\tilde{x})\ p_k(a'|\tilde{x})}.$$

- Compute predicted image label $a_{i,k}^*$:

$$a_{i,k}^* = \underset{a' \in \mathcal{A}}{\text{argmax}} \sum_{\tilde{x} \in \tilde{\mathcal{X}}} p_k(\tilde{x}|a)\ u(\tilde{x}, a'). \quad (17)$$

Return: Predicted image labels $\{a_{i,k}^*\}_{i,k=1}^{I,K}$.

[12] Abstractly, our interpretability algorithm performs a system identification of the neural network assumed to be a rationally inattentive agent. The proposed classification protocol in Algorithm 3 simply performs a function evaluation of the rational inattention model parameterized by the interpretability test outputs, namely, utility functions and sensing costs.

[13] We thank an anonymous reviewer for suggesting this idea.

design an auto-encoder for feature extraction, and replace the image class label with the image features as the state in the rational inattention model. This would result in a richer descriptive model of the CNN due to more degrees of freedom in the utility function. Such a framework facilitates us to study the performance of Algorithm 3. Also, by setting the loss function to be a combination of the prediction error (17) and reconstruction error (from the decoded image), one can train an auto-encoder to yield the *optimal* feature map that maximizes the prediction accuracy for the classification scheme of Algorithm 3.

Our proposed interpretable model generates a concave utility function by design. This is an important feature of the revealed preference framework; even though the actual deep learner's utility may not be convex. To quote Varian [16]: "If data can be rationalized by any non-trivial utility function, then it can be rationalized by a nice utility function. Violations of concavity cannot be detected with only a finite number of observations." A more speculative extension is to investigate the asymptotic behavior of the BRP and S-BRP decision tests for rationally inattentive utility maximization-do the tests pass when the number of deep CNNs tend to infinity? Recent results [50] show that an infinite dataset can at best be rationalized by a quasi-concave utility function.

Acknowledgement

This research was supported in part by the Army Research Office under grants W911NF-21-1-0093 and W911NF-24-1-0083, and the National Science Foundation under grants CCF-2112457 and CCF-2312198.

References

[1] C. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

[2] C. Sims. Rational inattention and monetary economics. *Handbook of Monetary Economics*, 3:155–181, 2010.

[3] M. Woodford. Inattentive valuation and reference-dependent choice. 2012.

[4] F. Matějka and A. McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

[5] H. De Oliveira, T. Denti, M. Mihm, and K. Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

[6] A. Caplin and D. Martin. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

[7] A. Caplin and M. Dean. Revealed preference, rational inattention, and costly information acquisition. *The American Economic Review*, 105(7):2183–2203, 2015.

[8] S. Nitinawarat and V. V. Veeravalli. Controlled sensing for sequential multihypothesis testing with controlled markovian observations and non-uniform control cost. *Sequential Analysis*, 34(1):1–24, 2015.

[9] V. Krishnamurthy and H. V. Poor. A tutorial on interactive sensing in social networks. *IEEE Transactions on Computational Social Systems*, 1(1):3–21, 2014.

[10] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

[11] D. Liang, J. Cheng, Z. Ke, and L. Ying. Deep magnetic resonance image reconstruction: Inverse problems meet neural networks. *IEEE Signal Processing Magazine*, 37(1):141–151, 2020.

[12] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[13] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[14] S. N. Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

[15] W. E. Diewert. Afriat and revealed preference theory. *The Review of Economic Studies*, 40(3):419–425, 1973.

[16] H. R. Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

[17] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6, 2017.

[18] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.

[20] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[24] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, 2016.

[25] P. Hase, C. Chen, O. Li, and C. Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7-1, pages 32–40, 2019.

[26] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

[27] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[28] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[30] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[31] H. Wang and D.-Y. Yeung. Towards Bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, 2016.

[32] P. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, 1981.

[33] L. Huang and H. Liu. Rational inattention and portfolio selection. *The Journal of Finance*, 62(4):1999–2040, 2007.

[34] S. E. Mirsadeghi, A. Royat, and H. Rezatofighi. Unsupervised image segmentation by mutual information maximization and adversarial regularization. *IEEE Robotics and Automation Letters*, 2021.

[35] W. Hoiles, V. Krishnamurthy, and K. Pattanayak. Rationally Inattentive Inverse Reinforcement Learning Explains YouTube commenting behavior. *The Journal of Machine Learning Research*, 21(170):1–39, 2020.

[36] K. Pattanayak and V. Krishnamurthy. Unifying classical and Bayesian revealed preference, 2021.

[37] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[38] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[42] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

[43] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006.

[44] S. N. Afriat. Efficiency estimation of production functions. *International economic review*, pages 568–598, 1972.

[45] R. Anand, T. Shanthi, R. Sabeenian, and S. Veni. Real time noisy dataset implementation of optical character identification using CNN. *International Journal of Intelligent Enterprise*, 7(1-3):67–80, 2020.

[46] Y. Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[50] P. J. Reny. A characterization of rationalizable consumer behavior. *Econometrica*, 83(1):175–192, 2015.

[51] D. Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[52] J. Gu, V. Tresp, and Y. Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pages 404–421. Springer, 2022.

[53] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik. Imagenet-21k pretraining for the masses. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.

[54] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[56] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[57] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

APPENDIX

A. Proof of Theorem 1

Proof of necessity of NIAS and NIAC:

1) NIAS (8): For agent $k \in \mathcal{K}$, define the subset $\mathcal{Y}_a \subseteq \mathcal{Y}$ so that for any observation $y \in \mathcal{Y}_a$, given posterior pmf $p_k(x|y)$, the optimal choice of action is a (3). We define the revealed posterior pmf given action a as $p_k(x|a)$. The revealed posterior pmf is a stochastically garbled version of the actual posterior pmf $p_k(x|y)$, that is,

$$p_k(x|a) = \sum_{y\in\mathcal{Y}} \frac{p_k(x,y,a)}{p_k(a)} = \sum_{y\in\mathcal{Y}} p_k(y|a)p_k(x|y) \quad (18)$$

Since the optimal action is a for all $y \in \mathcal{Y}_a$, (3) implies:

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \le 0 \\
\implies &\sum_{y\in\mathcal{Y}_a} p_k(y|a) \sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \le 0 \\
\implies &\sum_{y\in\mathcal{Y}} p_k(y|a) \sum_{x\in\mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \le 0 \\
&(\text{since } p_k(y|a) = 0,\ \forall y \in \mathcal{Y}\backslash\mathcal{Y}_a) \\
\implies &\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_k(y|a)p_k(x|y)(u_k(x,b) - u_k(x,a)) \le 0 \\
\implies &\sum_{x\in\mathcal{X}} p_k(x|a)(u_k(x,b) - u_k(x,a)) \le 0 \ (\text{from (18)})
\end{aligned}$$

This is precisely the NIAS inequality (8).

2) NIAC (9): Let $c_k = C(\alpha_k) > 0$, where $C(\cdot)$ denotes the information acquisition cost of the collection of agents $\mathcal{K}$. Also, let $J(\alpha_k, u_k)$ denote the expected utility of the k^{th} agent given attention strategy α_k (first term in RHS of (4)). Here, the expectation is taken wrt both the state x and observation y. It can be verified that $J(\cdot, u_k)$ is convex in the first argument. Finally, for the k^{th} agent, we define the revealed attention strategy α'_k over the set of actions $\mathcal{A}$ as $\alpha'_k(a|x) = p_k(a|x),\ \forall a \in \mathcal{A}$, where the variable $p_k(a|x)$ is obtained from the dataset $\mathbb{D}$. Clearly, the revealed attention strategy is a stochastically garbled version of the true attention strategy since

$$\alpha'_k(a|x) = p_k(a|x) = \sum_{y\in\mathcal{Y}} p_k(a|y)\alpha_k(y|x) \quad (19)$$

From Blackwell dominance [51] and the convexity of the expected utility functional $J(\cdot, u_k)$, it follows that:

$$J(\alpha'_k, u_j) \le J(\alpha_k, u_j), \quad (20)$$

when α_k Blackwell dominates α'_k. The above relationship holds with equality if $k = j$ (this is due to NIAS (8)). We now turn to condition (4) for optimality of attention strategy. The following inequalities hold for any pair of agents $j \neq k$:

$$\begin{aligned}
J(\alpha'_k, u_k) - c_k &\overset{(20)}{=} J(\alpha_k, u_k) - c_k \\
&\overset{(4)}{\ge} J(\alpha_j, u_k) - c_j \overset{(20)}{\ge} J(\alpha'_j, u_k) - c_j. \quad (21)
\end{aligned}$$

This is precisely the NIAC inequality (9).

Proof for sufficiency of NIAS and NIAC: Let $\{u_k, c_k\}_{k=1}^K$ denote a feasible solution to the NIAS and NIAC inequalities of Theorem 1. To prove sufficiency, we construct an UMRI tuple as a function of dataset $\mathbb{D}$ and the feasible solution that satisfies the optimality conditions (3),(4) of Definition 1.

Consider the following UMRI model tuple:

$$\begin{aligned}
&\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y} = \mathcal{A}, \mathcal{A}, \pi_0, C, \{p_k(a|x), u_k, k \in \mathcal{K}\}), \text{ where} \\
&C(p(a|x)) = \max_{k\in\mathcal{K}} c_k + J(p(a|x), u_k) - J(p_k(a|x), u_k).
\end{aligned} \quad (22)$$

In (22), $C(\cdot)$ is a convex cost since it is a point-wise maximum of monotone convex functions. Further, since NIAC is satisfied, (22) implies $C(\alpha_k) = c_k$. It only remains to show that

inequalities (3) and (4) in Definition 1 are satisfied for all agents in $\mathcal{K}$.

1) *NIAS implies* (3) *holds.* This is straightforward to show since the observation and action sets are identical.
2) *Information Acquisition Cost* (22) *implies* (4) *holds.* Fix agent $j \in \mathcal{K}$. Then, for any attention strategy $p(a|x)$, the following inequalities hold.

$$\begin{aligned} & C(p(a|x)) = \max_{k\in\mathcal{K}} \; c_k + J(p(a|x), u_k) - J(p_k(a|x), u_k) \\ \implies & J(p_j(a|x)) - c_j \geq J(p(a|x)) - C(p(a|x)), \; \forall \; p(a|x) \\ \implies & p_k(a|x) \in \underset{p(a|x)}{\operatorname{argmax}} J(p(a|x), u_k) - C(p(a|x)) = (4). \end{aligned}$$

B. S-UMRI (Sparse UMRI) Model for Rationally Inattentive Bayesian Utility Maximization

In Sec. II-A, we outlined the UMRI model for rationally inattentive utility maximization of K Bayesian agents parameterized by K utility functions and a cost of information acquisition. This section proposes a sparse version of the UMRI model, namely, the S-UMRI model that is parameterized by a *single* utility function that rationalizes the decisions of K Bayesian agents. Abstractly, the S-UMRI model is described by the tuple

$$\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, C, u, \{\alpha_k, \lambda_k, k \in \mathcal{K}\}). \tag{23}$$

All parameters in (23) are identical to that in (1) except for the additional parameter $\lambda_k \in \mathbb{R}_+$. λ_k can be interpreted as the sensitivity to information acquisition of the k^{th} agent. We discuss the significance of λ_k in more detail below. In complete analogy to Definition 1, Definition 3 below specifies the optimal action and attention strategy policy of the Bayesian agents $\mathcal{K}$.

Definition 3 (Rationally Inattentive Utility Maximization for S-UMRI)**.** *Consider a collection of Bayesian agents $\mathcal{K}$ parameterized by Θ in* (23) *under the S-UMRI model. Then,*
(a) **Expected Utility Maximization:** *Given posterior pmf $p(x|y)$, agent $k \in \mathcal{K}$ chooses action a that maximizes its expected utility:*

$$a \in \underset{a'\in\mathcal{A}}{\operatorname{argmax}} \; \mathbb{E}_x\{u_k(x, a')|y\} = \sum_{x\in\mathcal{X}} p(x|y)u(x, a') \tag{24}$$

(b) **Attention Strategy Rationality:** *Agent k chooses attention strategy α_k that optimally trades off between utility maximization and cost minimization.*

$$\alpha_k \in \underset{\alpha'}{\operatorname{argmax}} \, \mathbb{E}_y\{ \max_{a\in\mathcal{A}} \mathbb{E}_x\{u(x, a)|y\}\} - \lambda_k C(\alpha', \pi_0) \tag{25}$$

Remarks. 1. *Role of λ_k.* In (25), λ_k is the differentiating parameter across agents. Even though all agents have the same utility function, different values of λ_k result in different optimal strategies α_k (25).
2. *Sparsity of S-UMRI*. The UMRI and S-UMRI model tuples for K Bayesian agents are parameterized using $K(|\mathcal{X}||\mathcal{A}| + 1)$ and $|\mathcal{X}||\mathcal{A}| + K$ variables, respectively. The difference in variables for parametrization is linear in K.

Finally, in complete analogy to Theorem 1, we now state Theorem 3 that states necessary and sufficient conditions for the decisions of a collection of Bayesian agents to be rationalized by the S-UMRI model.

Theorem 3 (S-BRP Test for Rationally Inattentive Utility Maximization)**.** *Given the dataset $\mathbb{D}$* (5) *obtained from a collection of Bayesian agents $\mathcal{K}$. Then,*
1. *Existence: There exists a S-UMRI tuple $\Theta(\mathbb{D})$* (1) *that rationalizes dataset $\mathbb{D}$ if and only if there exists a feasible solution that satisfies the set of inequalities*

$$\textit{S-BRP}(\mathbb{D}) \leq \mathbf{0}. \tag{26}$$

In (6)*, S-BRP$(\cdot)$ corresponds to a set of inequalities stated in Algorithm 4 below. The set-valued estimate of Θ that rationalizes $\mathbb{D}$ is the set of all feasible solutions to* (6)*.*
2. *Reconstruction: Given a feasible solution $\{u, \lambda_k, c_k\}_{k=1}^K$ to $S - BRP(\mathbb{D}, \cdot)$, u is the k^{th} Bayesian agent's utility function, for all $k = 1, 2, \ldots, K$. The feasible cost of information acquisition C in $\Theta(\mathbb{D})$ is defined in terms of the feasible variables c_k, λ_k as:*

$$C(\alpha) = \max_{k\in\mathcal{K}} c_k + \lambda_k \sum_{x,a} (p(x, a) - p_k(x, a))u(x, a) \tag{27}$$

The proof of Theorem 3 closely follows the proof of Theorem 1 and hence, omitted. In comparison to the BRP test of Theorem 1, the S-BRP test has the same number of inequalities but fewer decision variables. Hence, the set of feasible parameters generated from Algorithm 4 is smaller compared to Algorithm 1.

Algorithm 4 S-BRP Convex Feasibility Test of Theorem 3

Require: Dataset $\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}\}$ from a collection of Bayesian agents $\mathcal{K}$.

Find: Positive reals $c_k, \lambda_k, u \in (0, 1]$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, $k \in \mathcal{K}$ that satisfy the following inequalities:

1. $\sum_x p_k(x|a) \; (u(x, b) - u(x, a)) \leq 0, \forall a, b, k,$
2. $\sum_{x,a} (p_j(x, a) - p_k(x, a))u(x, a) + \lambda_k(c_k - c_j) \leq 0, \forall j, k,$

where $p_k(x, a) = \pi_0(x)p_k(a|x), \; p_k(x|a) = \frac{p_k(x,a)}{\sum_{x'} p_k(x',a)}$.

Return: Set of feasible utility function u, scalars λ_k and information acquisition costs c_k incurred by agents $k \in \mathcal{K}$.

C. Construction of Deep CNN Dataset

We now explain how the decisions of the deep CNNs are incorporated into our main theorems Theorems 1 and 3. Suppose K deep CNNs indexed by $k = 1, \ldots, K$ with different training parameters are trained on the CIFAR-10 dataset. For every trained deep CNN k, given test image i from CIFAR-10 test dataset with image class s_i, let the vector $f_{i,k} \in \Delta^9$ denote the corresponding softmax output of the deep CNN. The vector $f_{i,k}$ is a 10-dimensional probability vector where $f_{i,k}(j)$ is the probability that deep CNN k classifies test image i as class j.

The decisions of all K deep CNNs on the CIFAR-10 test dataset are aggregated into dataset $\mathbb{D}$ for compatibility with Theorems 1 and 3 as follows:

$$\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \{1, 2, \ldots, K\}\}, \text{ where}$$
$$\pi_0(x) = \sum_{i=1}^{N} \frac{\mathbb{1}\{s_i = x\}}{N}, \; p_k(a|x) = \frac{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\} f_{i,k}(a)}{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\}},$$
$$N = 10^4, \; \mathcal{X} = \mathcal{A} = \{1, 2, \ldots 10\}. \tag{28}$$

Here $\pi_0(x)$ is the empirical probability that the image class of a test image in the CIFAR-10 test dataset is x. Since the output of the CNN is a probability vector, we compute $p_k(a|x)$ for the k^{th} CNN by averaging the a^{th} component of the output over all test images in image class x. Finally, N is the number of test images in the CIFAR-10 test dataset, and the set of true and predicted image classes are the same, i.e., $\mathcal{X} = \mathcal{A}$. Although implicit in the above description, our Bayesian revealed preference approach to interpretable deep image classification assumes the deep CNN's (agent's) ground truth is the true image label, and its decision a is the predicted image label.

D. Interpretable Deep Image Classification using Vision Transformer (ViT) Neural Network Trained on Large Image Datasets

In this appendix, we report two additional experiments to further illustrate the applicability of our approach to interpretable deep image classification in Sec. III. The experiments use the state-of-the-art neural network architecture, the vision transformer (ViT) [49], and are demonstrated on Tiny-Imagenet and CIFAR-100 datasets. The ViT is considered state-of-the-art in the field of computer vision due to its powerful self-attention mechanism. ViT models are reported to outperform CNNs in terms of computational efficiency and accuracy [52].

Dataset Construction: We consider the two datasets to demonstrate the applicability of our methodology on larger state spaces and show our result using $|\mathcal{X}| = 80$ states and $|\mathcal{A}| = 80$ actions. The datasets for the two experiments are constructed using:
(a) different noise levels in the training data, and (b) different training epochs. We describe the general training procedure used and then explain each experiment in detail. We perform a fine-training classification task on the respective datasets using the embeddings of a vision transformer pre-trained on the Imagenet-21k dataset [53]–[55].

Architecture: The vision transformer (ViT) is a neural network used for image classification and other computer vision tasks, representing images as a sequence of patches fed into a transformer encoder. It has shown promising results over convolutional neural networks in fundamental computer vision tasks. A fully connected linear layer is attached to the last layer of the ViT, which outputs the logit probabilities for classification. Like in Sec. III, the state and action denote the true and predicted label of an image, respectively. For both experiments, the feasible utilities are reconstructed using (a) the sparsity-enhanced UMRI model via (13), and (b) maximizing the margin of the BRP inequalities (15).

Experiment 1. ViT trained on Tiny ImageNet: In the first experiment, we run a classification task on 80 classes of the Tiny Imagenet dataset, which contains 100000 centered and cropped training images of 200 different classes [56]. We artificially add zero-mean Gaussian noise of 5 different values of noise variance ($\sigma^2 \in \{0.001, 0.004, 0.01, 0.04, 0.1\}$). We report the reconstructed utilities for the experiment using the sparsity enhanced UMRI model in Fig. 6 and by optimizing the robustness measure in Fig. 4. The cost of information acquisition is also reported against different epochs in Fig. 8(a). The robustness value for the inequalities (15) is computed as 0.01. In Table IV, we show the reconstructed robust utility functions computed via (15) for $|\mathcal{X}| = 25$ states and $|\mathcal{A}| = 25$ actions (total 3125 variables). Although (15) yields 5 utility functions, each of dimension 25x25, we only display the utility values for 1 trained neural network.

Experiment 2. ViT trained on CIFAR-100: In the second experiment, we run a classification task on 80 classes of the CIFAR-100 dataset, which contains 60000 centered and cropped images of 100 different artifacts [57]. We create the dataset by considering different training rounds as a decision problem. We run the fine-training task for 10 epochs and capture the confusion matrix at $\{2, 4, 6, 8, 10\}$ epochs to construct the interpretability dataset. We report the reconstructed utilities for the experiment using the sparsity enhanced UMRI model in Fig. 7 and by optimizing the robustness measure in Fig. 5. The cost of information acquisition is also reported against different variances in Fig 8(b). The robustness value for the inequalities (Def. 2) is computed as 0.02. Table V reports the max-margin utilities for both experiments computed via (15).

Computational Cost: The number of NIAS (8) and NIAC (9) inequalities combined are of the order $O(|\mathcal{K}|^2|\mathcal{A}|^2)$. The number of variables (utility values and cost of information acquisition) to be optimized for obtaining the robust utility and cost estimate (Definition 2) is of the order $O(|\mathcal{K}||\mathcal{A}|^2)$ (for the case when $|\mathcal{X}| = |\mathcal{A}|$). In the results shown below, we set the off-diagonal elements of the utility functions to 0 when computing sparse and robust point estimates of utility functions and costs for the trained neural networks to make the reconstruction computationally tractable. For completeness, we reconstruct the robust point estimate for ViT for $|\mathcal{X}| = 25$ and $|\mathcal{K}| = 5$ without assuming the utility function is a diagonal matrix; the results are displayed in Table IV.

Supplementary Document

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes Figures 4, 5, 6, 7, 8(a), and 8(b) and Tables V and IV. Contact Kunal Pattanayak and Adit Jain for further questions about this work.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

Vikram Krishnamurthy (F'05) received the Ph.D. degree from the Australian National University in 1992. He is a professor in the School of Electrical & Computer Engineering, Cornell University. From 2002-2016 he was a Professor and Canada Research Chair at the University of British Columbia, Canada. His research interests include statistical signal processing and stochastic control in social networks and adaptive sensing. He served as Distinguished Lecturer for the IEEE Signal Processing Society and Editor-in-Chief of the IEEE Journal on Selected Topics in Signal Processing. In 2013, he was awarded an Honorary Doctorate from KTH (Royal Institute of Technology), Sweden. He is author of two books *Partially Observed Markov Decision Processes* and *Dynamics of Engineered Artificial Membranes and Biosensors* published by Cambridge University Press in 2016 and 2018, respectively.

Kunal Pattanayak (S'21) received the integrated Bachelors and Masters in Technology degrees in Electronics and Electrical Communication Engineering from Indian Institute of Technology, Kharagpur in 2018 and Ph .D degree in Electrical and Computer Engineering from Cornell University, USA in 2023. His research interests include inverse reinforcement learning, behavioral economics, statistical signal processing and design of counter-autonomous systems. He is a recipient of the McMullen graduate fellowship by Cornell University, and has been a speaker at the 2020 Sloan-NOMIS Conference on Attention and Applied Economics. He is currently an associate in Liquidity Risk at Goldman Sachs, USA.

Adit Jain Adit Jain received his Bachelor of Technology in Electronics and Communication Engineering from the Indian Institute of Technology, Guwahati in 2022. He is currently a graduate student at Cornell University. His research interests include structural results for episodic reinforcement learning, non-convex optimization, and high-dimensional linear bandits. His research is focused on applications in federated learning and large language models. He is a recipient of the Data Science Fellowship by the Cornell Center for Social Sciences.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

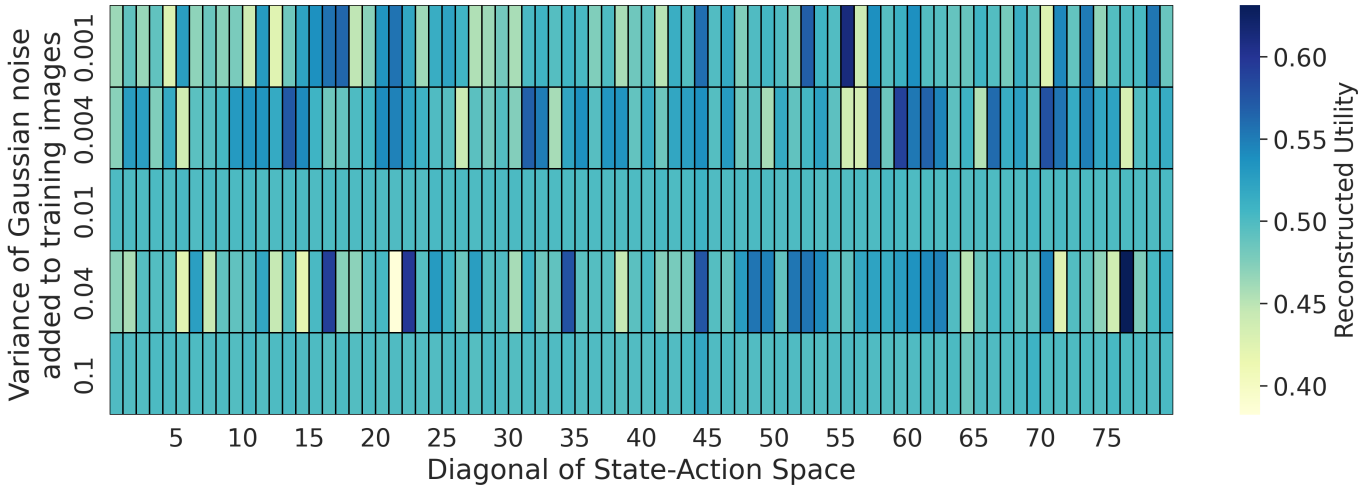


Fig. 4: Interpretable robust diagonal utility values ($\{u_k(x,x),\ x \in \mathcal{X}, k \in \mathcal{K}\}$) for UMRI for $K = 5$ and $|\mathcal{X}| = 80$ states, which maximize the robustness value in the UMRI model (Def. 2). The dataset is constructed using confusion matrices aggregated by training the vision transformer on Tiny-Imagenet dataset with additive Gaussian noise, with varying values of noise variance (Experiment 1 in Appendix D). The utility values are normalized and lie in the interval [0,1].

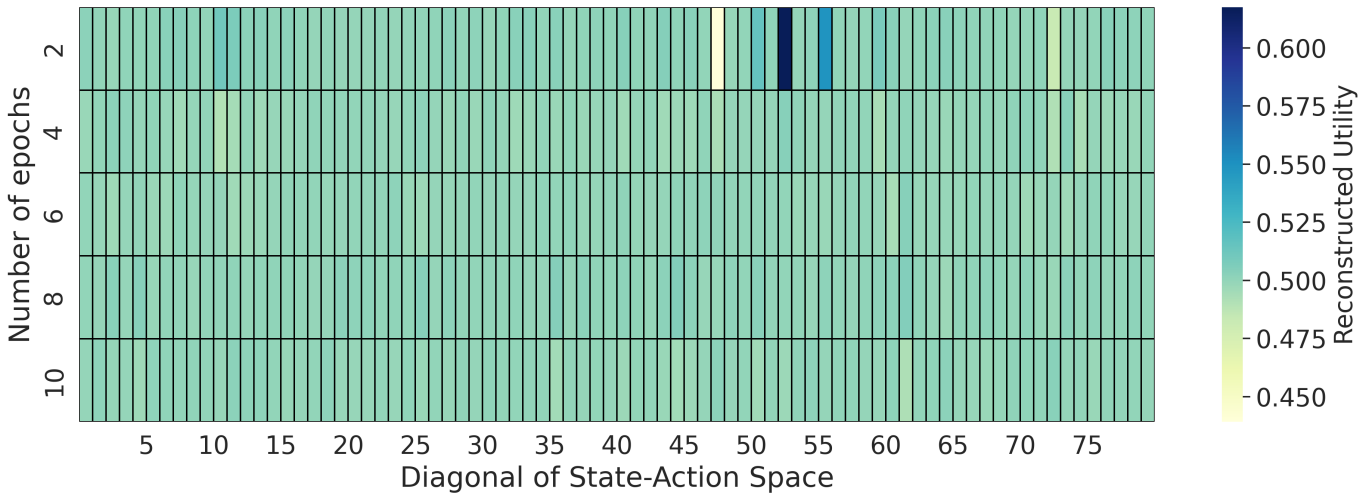


Fig. 5: Interpretable robust diagonal utility values ($\{u_k(x,x),\ x \in \mathcal{X}, k \in \mathcal{K}\}$) for UMRI for $|\mathcal{K}| = 5$ and $\mathcal{X}| = 80$ states, which maximize the robustness value in the UMRI model (Def. 2). The dataset is constructed using confusion matrices aggregated at different epochs while training the vision transformer on the CIFAR-100 dataset (Experiment 2 in Appendix D). The utility values are normalized and lie in the interval [0,1].

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

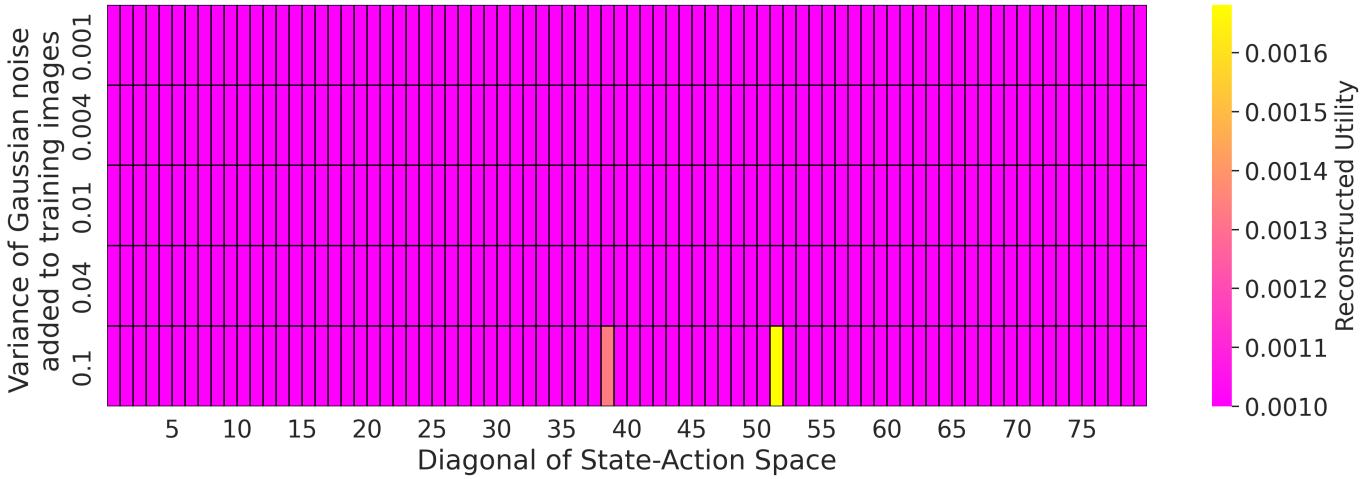


Fig. 6: Interpretable diagonal utility values ($\{u_k(x,x),\ x \in \mathcal{X}, k \in \mathcal{K}\}$) for UMRI for $|\mathcal{K}| = 5$ and $|\mathcal{X}| = 80$ states, computed using sparsity enhanced UMRI model (Theorem 2). The dataset is constructed using confusion matrices aggregated by training the vision transformer on Tiny-Imagenet dataset with additive Gaussian noise, with varying noise variances (Experiment 1 in Appendix D).

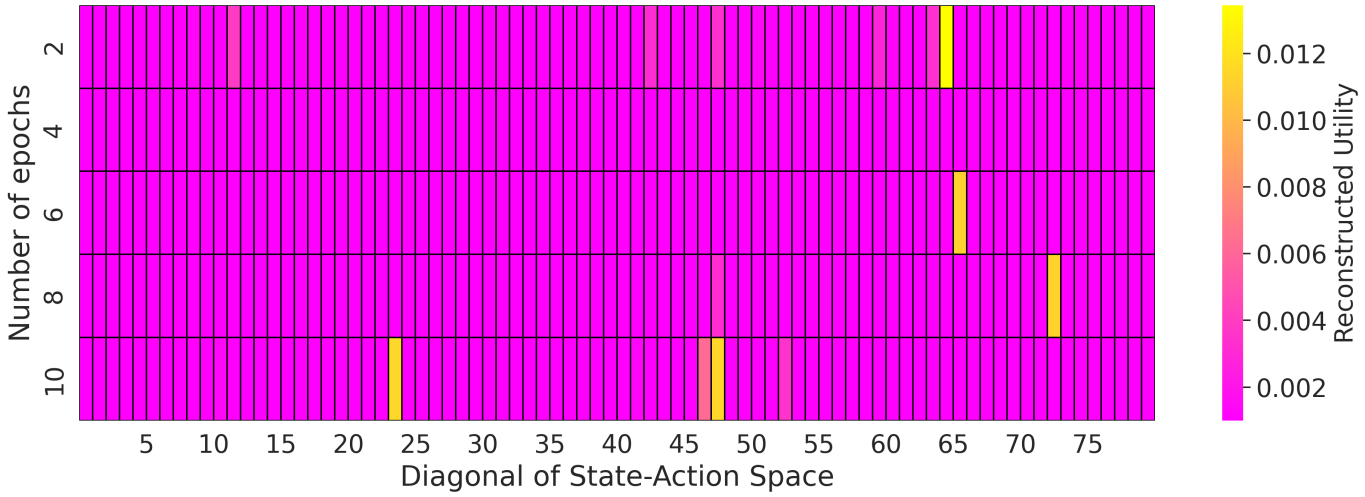
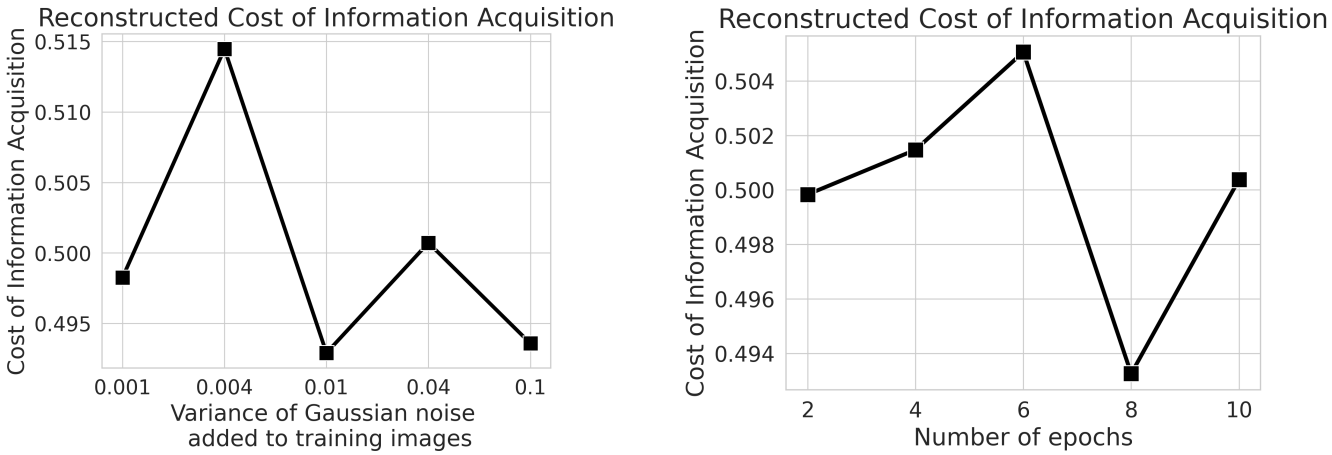


Fig. 7: Interpretable diagonal utility values ($\{u_k(x,x),\ x \in \mathcal{X}, k \in \mathcal{K}\}$) for UMRI for $K = 5$ and $|\mathcal{X}| = 80$ states, computed using sparsity enhanced UMRI model (Theorem 2). The dataset is constructed using confusion matrices aggregated at different epochs while training the vision transformer on the CIFAR-100 dataset (Experiment 2 in Appendix D).

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335



(a) Reconstructed cost for Experiment 1

(b) Reconstructed cost for Experiment 2

Fig. 8: Cost of information acquisition C reconstructed in Experiment-1 and Experiment-2 (Appendix D) corresponding to the robustness value $\mathcal{R}_{\text{BRP}}(\mathbb{D})$ for the UMRIThe utility values in the table above are normalized and lie in the interval [0,1].

	Action Space																							
State Space	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0	0.99	0.49	0.52	0.49	0.49	0.49	0.48	0.49	0.49	0.48	0.43	0.46	0.49	0.49	0.48	0.48	0.48	0.49	0.49	0.47	0.47	0.48	0.48	0.49	0.48
1	0.48	0.93	0.49	0.49	0.48	0.49	0.48	0.49	0.49	0.48	0.43	0.48	0.43	0.49	0.48	0.48	0.48	0.49	0.48	0.48	0.47	0.48	0.47	0.49	0.49
2	0.42	0.49	1.0	0.49	0.4	0.48	0.48	0.48	0.47	0.48	0.45	0.42	0.49	0.42	0.48	0.48	0.48	0.49	0.49	0.49	0.48	0.48	0.48	0.41	0.49
3	0.48	0.49	0.49	1.0	0.46	0.49	0.48	0.49	0.47	0.48	0.43	0.48	0.49	0.41	0.48	0.48	0.48	0.49	0.49	0.49	0.45	0.49	0.49	0.47	0.48
4	0.48	0.46	0.49	0.46	1.0	0.42	0.48	0.49	0.46	0.48	0.37	0.47	0.49	0.49	0.48	0.48	0.48	0.49	0.49	0.49	0.48	0.49	0.48	0.49	0.48
5	0.48	0.49	0.49	0.42	0.49	0.96	0.48	0.49	0.49	0.48	0.43	0.48	0.49	0.49	0.48	0.47	0.48	0.49	0.41	0.49	0.47	0.48	0.48	0.49	0.48
6	0.48	0.49	0.49	0.47	0.49	0.49	0.94	0.49	0.49	0.48	0.43	0.48	0.49	0.49	0.48	0.48	0.48	0.49	0.49	0.48	0.47	0.48	0.48	0.49	0.48
7	0.48	0.49	0.49	0.49	0.49	0.49	0.48	0.96	0.49	0.48	0.44	0.44	0.49	0.49	0.48	0.48	0.48	0.49	0.49	0.48	0.47	0.49	0.48	0.49	0.47
8	0.48	0.49	0.49	0.49	0.49	0.49	0.48	0.49	0.95	0.48	0.43	0.48	0.49	0.49	0.49	0.48	0.48	0.44	0.49	0.48	0.47	0.48	0.48	0.49	0.48
9	0.48	0.49	0.5	0.49	0.49	0.49	0.48	0.49	0.49	0.93	0.47	0.3	0.49	0.49	0.48	0.49	0.48	0.43	0.49	0.48	0.47	0.48	0.48	0.49	0.48
10	0.49	0.49	0.35	0.48	0.55	0.49	0.48	0.4	0.49	0.48	1.0	0.51	0.46	0.49	0.48	0.48	0.49	0.49	0.49	0.48	0.41	0.48	0.5	0.49	0.48
11	0.46	0.49	0.49	0.49	0.48	0.48	0.48	0.53	0.49	0.48	0.29	1.0	0.49	0.49	0.48	0.48	0.48	0.43	0.49	0.49	0.49	0.48	0.48	0.49	0.48
12	0.48	0.47	0.49	0.49	0.45	0.49	0.48	0.49	0.49	0.48	0.43	0.48	0.95	0.49	0.48	0.48	0.48	0.49	0.49	0.48	0.47	0.48	0.48	0.49	0.48
13	0.48	0.49	0.49	0.49	0.48	0.49	0.48	0.49	0.49	0.48	0.43	0.48	0.49	0.94	0.48	0.48	0.48	0.49	0.41	0.49	0.47	0.48	0.48	0.49	0.48
14	0.48	0.49	0.49	0.42	0.49	0.49	0.48	0.49	0.46	0.48	0.43	0.48	0.49	0.49	0.92	0.48	0.48	0.49	0.48	0.47	0.47	0.48	0.48	0.49	0.48
15	0.48	0.49	0.49	0.48	0.49	0.5	0.48	0.49	0.49	0.48	0.43	0.48	0.49	0.49	0.48	0.98	0.48	0.5	0.49	0.48	0.47	0.48	0.48	0.49	0.48
16	0.48	0.49	0.5	0.49	0.39	0.48	0.48	0.49	0.49	0.48	0.47	0.34	0.49	0.47	0.48	0.48	0.93	0.49	0.49	0.48	0.47	0.48	0.48	0.49	0.48
17	0.48	0.49	0.5	0.49	0.49	0.49	0.48	0.49	0.49	0.48	0.33	0.33	0.49	0.49	0.48	0.46	0.49	0.95	0.49	0.48	0.49	0.48	0.48	0.48	0.48
18	0.48	0.49	0.49	0.49	0.49	0.49	0.48	0.48	0.49	0.48	0.35	0.48	0.49	0.52	0.49	0.48	0.49	0.49	1.0	0.44	0.48	0.48	0.48	0.49	0.48
19	0.49	0.48	0.46	0.48	0.48	0.49	0.48	0.46	0.49	0.48	0.43	0.48	0.49	0.45	0.49	0.47	0.48	0.49	0.5	0.96	0.47	0.48	0.48	0.49	0.48
20	0.48	0.49	0.5	0.49	0.45	0.47	0.48	0.49	0.49	0.48	0.46	0.37	0.49	0.49	0.48	0.48	0.46	0.47	0.43	0.49	1.0	0.48	0.48	0.52	0.48
21	0.49	0.48	0.42	0.46	0.48	0.43	0.48	0.48	0.49	0.48	0.43	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.49	0.48	0.47	0.92	0.48	0.49	0.48
22	0.48	0.49	0.49	0.48	0.49	0.49	0.48	0.49	0.48	0.48	0.34	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.49	0.48	0.48	0.48	0.93	0.41	0.48
23	0.48	0.49	0.49	0.5	0.49	0.49	0.48	0.47	0.49	0.48	0.43	0.48	0.49	0.43	0.48	0.48	0.48	0.49	0.5	0.48	0.41	0.48	0.48	0.96	0.48
24	0.49	0.49	0.43	0.49	0.48	0.49	0.48	0.47	0.49	0.48	0.43	0.48	0.45	0.49	0.48	0.49	0.48	0.49	0.48	0.48	0.47	0.48	0.48	0.49	0.94

TABLE IV: Interpretable utility values for a neural network trained on 0.001 noise variance ($\{u_0(x,a),\ x \in \mathcal{X}, a \in \mathcal{A}\}$) for UMRI for $K = 5$ and $X = 25$, for experiments 1 described in Appendix D. Although the robust utility estimates comprise 3125 variables, we only display the utilities for the first trained neural network. The utility values are normalized and lie in the interval [0,1]. Recall that the dataset $\mathcal{D}$ used to compute the above utility values are obtained from the vision transformer (ViT) by varying values of training image dataset (Tiny-Imagenet) noise variance in experiment 1. We observe that the utility values are largest along the diagonal. This is expected since the classification accuracy is close to 100%, hence the reconstructed UMRI is expected to have the largest incentive to choose the action $a = x$, where x denotes the true label of an image.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3381335

State Action Space	Trained NNs (1-5) for Experiment 1					Trained NNs (1-5) for Experiment 2				
0	0.46	0.47	0.5	0.47	0.5	0.5	0.5	0.5	0.5	0.5
1	0.49	0.53	0.5	0.46	0.5	0.5	0.5	0.5	0.5	0.5
2	0.47	0.53	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3	0.49	0.48	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
4	0.43	0.52	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
5	0.53	0.44	0.5	0.42	0.5	0.5	0.5	0.5	0.5	0.5
6	0.47	0.49	0.5	0.53	0.5	0.5	0.5	0.5	0.5	0.5
7	0.49	0.5	0.5	0.44	0.5	0.5	0.5	0.5	0.5	0.5
8	0.47	0.5	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
9	0.48	0.53	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
10	0.44	0.54	0.5	0.5	0.5	0.51	0.49	0.5	0.5	0.5
11	0.53	0.53	0.5	0.52	0.5	0.51	0.49	0.5	0.5	0.5
12	0.42	0.52	0.5	0.44	0.5	0.5	0.5	0.5	0.5	0.5
13	0.48	0.57	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
14	0.52	0.54	0.5	0.42	0.5	0.5	0.5	0.5	0.5	0.5
15	0.54	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
16	0.56	0.48	0.5	0.59	0.5	0.5	0.5	0.5	0.5	0.5
17	0.56	0.49	0.5	0.47	0.5	0.5	0.5	0.5	0.5	0.5
18	0.45	0.5	0.5	0.47	0.5	0.5	0.5	0.5	0.5	0.5
19	0.47	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
20	0.53	0.54	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
21	0.56	0.55	0.5	0.38	0.5	0.5	0.5	0.5	0.5	0.5
22	0.52	0.52	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5
23	0.46	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
24	0.51	0.5	0.5	0.53	0.5	0.5	0.5	0.5	0.5	0.5
25	0.52	0.5	0.5	0.51	0.5	0.5	0.5	0.5	0.5	0.5
26	0.51	0.44	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
27	0.45	0.49	0.5	0.53	0.5	0.5	0.5	0.5	0.5	0.5
28	0.46	0.48	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
29	0.48	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
30	0.46	0.47	0.5	0.46	0.5	0.5	0.5	0.5	0.5	0.5
31	0.5	0.57	0.5	0.51	0.5	0.5	0.5	0.5	0.5	0.5
32	0.51	0.55	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
33	0.5	0.46	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
34	0.51	0.52	0.5	0.58	0.5	0.5	0.5	0.5	0.5	0.5
35	0.5	0.54	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
36	0.48	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
37	0.5	0.53	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
38	0.46	0.53	0.5	0.44	0.5	0.5	0.5	0.5	0.5	0.5
39	0.48	0.49	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
40	0.49	0.52	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
41	0.45	0.48	0.5	0.47	0.5	0.5	0.5	0.5	0.5	0.5
42	0.51	0.52	0.5	0.48	0.5	0.5	0.5	0.5	0.5	0.5
43	0.5	0.53	0.5	0.48	0.5	0.5	0.5	0.5	0.5	0.5
44	0.56	0.54	0.51	0.58	0.52	0.5	0.5	0.5	0.51	0.5
45	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
46	0.52	0.53	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
47	0.48	0.48	0.5	0.53	0.5	0.44	0.49	0.5	0.5	0.5
48	0.5	0.5	0.5	0.56	0.5	0.5	0.5	0.5	0.5	0.5
49	0.51	0.46	0.5	0.55	0.5	0.5	0.5	0.5	0.5	0.5
50	0.5	0.52	0.5	0.49	0.5	0.52	0.5	0.5	0.5	0.5
51	0.48	0.49	0.5	0.55	0.5	0.5	0.5	0.5	0.5	0.5
52	0.57	0.51	0.5	0.55	0.5	0.62	0.5	0.5	0.5	0.5
53	0.51	0.53	0.5	0.54	0.5	0.5	0.5	0.5	0.5	0.5
54	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
55	0.61	0.44	0.5	0.49	0.5	0.55	0.5	0.5	0.5	0.5
56	0.44	0.43	0.5	0.52	0.5	0.5	0.5	0.5	0.5	0.5
57	0.54	0.57	0.5	0.52	0.5	0.5	0.5	0.5	0.5	0.5
58	0.5	0.49	0.5	0.51	0.5	0.5	0.5	0.5	0.5	0.5
59	0.51	0.59	0.5	0.53	0.5	0.51	0.49	0.5	0.5	0.5
60	0.54	0.55	0.5	0.54	0.5	0.5	0.5	0.49	0.5	0.5
61	0.5	0.57	0.5	0.54	0.5	0.5	0.5	0.5	0.51	0.49
62	0.5	0.55	0.5	0.55	0.5	0.5	0.5	0.5	0.5	0.5
63	0.49	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
64	0.49	0.51	0.5	0.45	0.49	0.5	0.5	0.5	0.5	0.5
65	0.51	0.45	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
66	0.5	0.56	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
67	0.48	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
68	0.51	0.53	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
69	0.49	0.49	0.5	0.49	0.51	0.5	0.5	0.5	0.5	0.5
70	0.43	0.58	0.5	0.55	0.5	0.5	0.5	0.5	0.5	0.5
71	0.54	0.55	0.5	0.43	0.5	0.5	0.5	0.5	0.5	0.5
72	0.49	0.52	0.5	0.5	0.5	0.48	0.49	0.5	0.5	0.5
73	0.55	0.55	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5
74	0.47	0.52	0.5	0.47	0.5	0.5	0.49	0.5	0.5	0.5
75	0.5	0.52	0.5	0.44	0.5	0.5	0.5	0.5	0.5	0.5
76	0.5	0.43	0.5	0.63	0.5	0.5	0.5	0.5	0.5	0.5
77	0.5	0.5	0.5	0.48	0.5	0.5	0.5	0.5	0.5	0.5
78	0.55	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
79	0.49	0.51	0.5	0.51	0.5	0.5	0.5	0.5	0.5	0.5

TABLE V: Interpretable robust diagonal utility values ($\{u_k(x,x),\ x \in \mathcal{X}, k \in \mathcal{K}\}$) for UMRI for $|\mathcal{K}| = 5$ and $|\mathcal{X}| = 80$, for experiments 1 and 2 described in Appendix D. Recall that the dataset $\mathcal{D}$ used to compute above utility values are obtained from vision transformer (ViT) network trained on varying values of training image dataset noise variance in experiment 1, and varying epochs in experiment 2, respectively. The utility values are normalized and lie in the interval [0,1].