ElementaryNet: A Non-Strategic Neural Network for Predicting Human Behavior in Normal-Form Games

Greg d'Eon[1], **Hala Murad**[1], **Kevin Leyton-Brown**[1], **James R. Wright**[2],
[1]University of British Columbia
[2]University of Alberta
gregdeon@cs.ubc.ca, hmurad01@student.ubc.ca, kevinlb@cs.ubc.ca, james.wright@ualberta.ca,

Abstract

Behavioral game theory models serve two purposes: yielding insights into how human decision-making works, and predicting how people would behave in novel strategic settings. A system called GameNet represents the state of the art for predicting human behavior in the setting of unrepeated simultaneous-move games, combining a simple "level-k" model of strategic reasoning with a complex neural network model of non-strategic "level-0" behavior. Although this reliance on well-established ideas from cognitive science ought to make GameNet interpretable, the flexibility of its level-0 model raises the possibility that it is able to emulate strategic reasoning. In this work, we prove that GameNet's level-0 model is indeed too general. We then introduce ElementaryNet, a novel neural network that is provably incapable of expressing strategic behavior. We show that these additional restrictions are empirically harmless, with ElementaryNet and GameNet having statistically indistinguishable performance. We then show how it is possible to derive insights about human behavior by varying ElementaryNet's features and interpreting its parameters, finding evidence of iterative reasoning, learning about the depth of this reasoning process, and showing the value of a rich level-0 specification.

Code & data — https://github.com/gregdeon/elementarynet
Extended version — https://arxiv.org/abs/2503.05925

1 Introduction

Human behavior in strategic settings often deviates significantly from the predictions of classical game theory: for instance, humans often play dominated actions (as in the unrepeated Prisoner's Dilemma), which a fully rational agent would not. Behavioral game theory aims to address these shortcomings by developing predictive models of human strategic behavior. The purpose of such models is two-fold. First, such models that have been trained on experimental data can be analyzed to yield insight about human psychology. In this sense, classical game theory can be seen as a zero-parameter model that predicts poorly; the goal here is to find models that make much more accurate predictions, but are as simple and interpretable as possible. Second, we might simply build models that are as accurate as possible. Such models are useful for developing agents that will interact with humans or for designing systems of rules ("mechanisms") that will perform well when faced with human participants.

Past work has made substantial progress towards the first goal, demonstrating that so-called iterative reasoning models with just two parameters are surprisingly good at predicting behavior in new strategic settings, and hence giving insight into the way human subjects reason. Let us describe Quantal Cognitive Hierarchy (QCH), the best performing such model (Camerer, Ho, and Chong 2004). QCH predicts that people perform strategic reasoning at some finite "level". The model includes a probability distribution describing the proportion of agents at each level; when this distribution is Poisson (having one parameter), we call the model QCHp. Level-0 reasoners are nonstrategic: they perform some arbitrary computation that falls short of forming beliefs about their opponents and best responding to these beliefs. Camerer, Ho, and Chong (2004) simply assert that level-0 reasoners choose actions uniformly at random, which is clearly nonstrategic (and parameter free); we call this instantiation of the model Uniform + QCHp. Level-1 reasoners quantally best respond to level-0 agents, where quantal response is a noisy version of best response that depends on a "precision" parameter. Level-k reasoners for $k \geq 2$ are more complex: they know the distribution over levels $0, \ldots, k-1$, and quantally best respond to the aggregate distribution of play by all lower-level agents. More recent work showed that even better performance can be achieved by hand-crafting a richer level-0 model using heuristics based on insights from cognitive psychology, at the expense of adding more parameters (Wright and Leyton-Brown 2019).

Regarding the second goal of raw predictive performance, it is perhaps unsurprising that the current state-of-the-art model for predicting human behavior in unrepeated normal-form games, dubbed GameNet (Hartford, Wright, and Leyton-Brown 2016), is based on a neural network. GameNet extends the Uniform + QCHp architecture by replacing the uniform level-0 specification with a learned model and the Poisson distribution by a finite-depth histogram. The learned model uses a permutation-equivariant architecture to express the inductive bias that an action should be played with the same probability regardless of its index in the game matrix, but is otherwise unrestricted. Empirically, GameNet dramatically outperforms Uniform + QCHp.

Although its neural network component is an uninterpretable black box, GameNet was based on the QCH architecture in part to allow for meaningful interpretation: one can ask about the shape of the level distribution, the precision parameter, and the way model performance varies when levels are added or removed. In this vein, Hartford, Wright, and Leyton-Brown (2016) made a striking observation: GameNet performed best when it was restricted to predict that all agents were level-0. What should we make of this result? Does it show that iterative strategic reasoning is a dead end, and that human subjects are better described in some different way? Or, does it show that a sufficiently general neural network is already able to simulate iterative strategic reasoning, and hence that GameNet's apparent interpretability was a mirage? Finally, how would we tell the difference?

In this work, we make three main contributions. First, in Section 3, we show that GameNet's purportedly level-0 neural network *is* capable of strategic reasoning, and is hence inappropriate to use for describing level-0 agents. Our proof is constructive: we give a specific setting of its parameters that computes quantal best response to maxmax, a strategic model. This finding helps to explain why adding levels of strategic reasoning did not improve GameNet's performance.

Second, in Section 4, we introduce *ElementaryNet*, a new restriction of GameNet's architecture that is only capable of non-strategic behavior. Our architecture is rooted in the concept of elementary models (Wright and Leyton-Brown 2022), a set of behavioral models which are provably incapable of representing strategic behavior. Elementary models compute a real-valued "potential" for each outcome of the game that summarizes the set of utilities for all the players into a single number, and then compute a distribution of play using only those potentials. Intuitively, this restriction to a single summary per outcome prevents the model from forming a belief about the opponent's play (based on the opponent's utilities), and then best responding to that belief (based on the model's own utilities). Our architecture is a convex combination of neural networks whose inputs are summarized by a single potential function each; it is thus a convex combination of elementary models, which is provably non-strategic in a precise, formal sense.

Finally, in Section 5, we perform extensive experiments on ElementaryNet. We first show that, despite the additional restrictions on its architecture, ElementaryNet combined with an iterative reasoning model achieves predictive power statistically indistinguishable from that of GameNet. We then show how the model's clean delineation between strategic and non-strategic components makes it possible to obtain interpretable insights about human behavior by varying different features of the model. In particular, we show that ElementaryNet performs significantly worse without a strategic model, demonstrating that iterative reasoning is indeed a good model of human behavior. We show that restricting the level-0 model—to only consider the player's own payoffs, or to use a fixed set of four basis potentials—degrades performance, demonstrating the value of a rich level-0 specification. Finally, we successfully interpret the parameters of the iterative reasoning model that are co-learned with ElementaryNet, in exactly the same way that was uninformative with GameNet.

2 Preliminaries

We begin by fixing notation. A *2-player* $n \times m$ *normal-form game* is a tuple $G = (A, u)$, where $A = A_1 \times A_2$ is the set of *action profiles*; $A_1 = \{1, \ldots, n\}$ and $A_2 = \{1, ..., m\}$ are the sets of *actions* available to agents 1 and 2, respectively; and $u = (u_1, u_2)$ is profile of *utility functions* $u_i : A \to \mathbb{R}$, each mapping an action profile to a real-valued utility for agent i. For convenience, we will sometimes refer to the utility matrices $U^1 = [u_1(a_i, a_j)]_{ij}$ and $U^2 = [u_2(a_i, a_j)]_{ij}$.

A *behavior* s_i is an element of the simplex $\Delta(A_i)$, representing a distribution over agent i's actions; we use the non-standard term "behavior", rather than the more standard "strategy", to avoid other awkward terminology, such as a "non-strategic strategy". A *behavior profile* is a tuple of behaviors $s = (s_1, s_2)$. Overloading notation, we denote an agent's expected utility as $u_i(s) = \mathbb{E}_{a \sim s} u_i(a)$. For either agent i, we write s_{-i} to represent the behavior of the other agent, and (s_i, s_{-i}) to refer to a behavior profile. A *behavioral model* is a function f_i that maps a game $G = (A, u)$ to a behavior $f_i(G) \in \Delta(A_i)$.

Existing Behavioral Models

We now describe several behavioral models used in prior behavioral game theory work. A common building block in many of these models is the concept of quantal best response.

Definition 2.1. *The (logit)* quantal best response *to a strategy* s_{-i} *is* $QBR_i(s_{-i}; \lambda, G)(a_i) \propto \exp[\lambda \cdot u_i(a_i, s_{-i})]$, *where* λ *(the* precision *parameter) controls the agent's sensitivity to differences in utilities.*

The quantal cognitive hierarchy (QCH) model (e.g., Wright and Leyton-Brown 2017) combines quantal best response with a model of iterative reasoning.

Definition 2.2. *Let* G *be a game,* $\lambda \in \mathbb{R}$ *be a precision,* s^0 *be a profile of level-0 behaviors in* G*, and* D *be a probability distribution over levels. Then, the* level-k quantal hierarchical behavior *is defined as* $s_i^k = QBR_i(s_{-i}^{0:k-1}; \lambda, G)$, *where* $s_i^{0:k-1}(a_i) \propto \sum_{m=0}^{k-1} D(m) s_i^m(a_i)$. *The* quantal cognitive hierarchy *behavior is the weighted average of these behaviors* $QCH_i(a_i) = \sum_k D(k) s_i^k(a_i)$.

QCH models depend critically on the level-0 model, which determines not just the behavior of level-0 agents, but also of higher-level agents who react to it. A natural, simple choice is the uniform distribution, which was originally used in the Level-k (Nagel 1995; Costa-Gomes, Crawford, and Broseta 2001) and Cognitive Hierarchy (Camerer, Ho, and Chong 2004) models. However, Wright and Leyton-Brown (2019) found improvements using richer level-0 models based on heuristics from cognitive science. For example, one such heuristic is the *maxmax behavioral model*,

$$f_i^{\text{maxmax}}(a_i) \propto \begin{cases} 1, & a_i \in \arg\max_{a_i'} \max_{a_{-i}} u_i(a_i', a_{-i}); \\ 0, & \text{otherwise.} \end{cases}$$

Notably, this heuristic depends only on the agent's utility u_i, neglecting the opponent's utility u_{-i}. The other heuristics also depend on simple linear combinations of the players' utilities, such as the sum (or *welfare*) $u_1 + u_2$ or the difference (*unfairness*) $u_1 - u_2$.

GameNet

GameNet (Hartford, Wright, and Leyton-Brown 2016) is a deep learning architecture for predicting human strategic behavior. It can be understood as taking the insight that a richer level-0 specification can yield better performance to its logical limit. Broadly, GameNet consists of two parts: *feature layers*, which are intended to model level-0 behavior, and *action response (AR) layers*, which perform iterative strategic reasoning. AR layers are essentially a generalization of the QCH model, with additional parameters independently controlling each level's precision, the distribution over lower levels to which agents respond, and the transformed utility matrices which they use for this response. We devote more attention to the feature layers.

Let the 0th hidden layer consist of the two matrices $H^{0,1} = U^1$ and $H^{0,2} = U^2$. For each hidden layer $0 \leq \ell < L$, the matrices are first transformed by *pooling units* into

$$P^{\ell,c} = \begin{cases} H^{\ell,c} & \text{if } c \leq C_\ell; \\ \text{rowmax}(H^{\ell,c-C_\ell}) & \text{if } C_\ell < c \leq 2C_\ell; \\ \text{colmax}(H^{\ell,c-2C_\ell}) & \text{if } 2C_\ell < c \leq 3C_\ell, \end{cases}$$

where rowmax and colmax are functions that replace each entry of a matrix with the maximum value in its row or column, respectively; that is, for all $X \in \mathbb{R}^{n\times m}$, $\text{rowmax}(X)_{ij} = \max_{1\leq a\leq m} X_{ia}$ and $\text{colmax}(X)_{ij} = \max_{1\leq a\leq n} X_{aj}$. Then, the next layer's *hidden units* are $H^{\ell,c} = \text{relu}\left(\sum_{c'=1}^{3C_{\ell-1}} w^\ell_{c,c'} P^{\ell-1,c'} + b^\ell_c\right)$, where $\{C_\ell\}_{0:L}$ describe the sizes of each hidden layer, including the input layer with size $C_0 = 2$, and the ReLU operation $\text{relu}(x) = \max\{0, x\}$ is applied pointwise.

After the final hidden layer, the matrices $\{H^{L,c}\}_{c=1:C_L}$ are transformed into a single distribution over the row player's actions as $f = \sum_{c=1}^{C_L} w_c f^c$, where $f^c = \text{softmax}(\sum_j H^{L,c}_{i,j})$, $\text{softmax}(x)_i = \exp(x_i)/\sum_{i'} \exp(x_{i'})$, and the weights $w_c \in \Delta(C_L)$ are subject to simplex constraints. An analogous predicted distribution over the column player's actions is made by replacing the input utility matrices U^1 and U^2 with $(U^2)^T$ and $(U^1)^T$, respectively. This architecture is summarized in Figure 1.

Compared to a more standard, off-the-shelf architecture—e.g., flattening the utility matrices into a vector of length $2nm$ and applying a feedforward neural network—feature layers have several advantages. One is that they are *permutation equivariant*, which guarantees that permuting the utility matrices will permute the predictions correspondingly. Permutation equivariance lowers the number of parameters of the network, making it learn more efficiently and removing the need for data augmentation. Feature layers are also agnostic to the size of the game, making it possible to learn from heterogeneous data with a variety of action spaces and to generalize to games of new sizes.

Non-Strategic Behavioral Models

To make precise claims about whether or not a model is strategic, we adopt a formal definition of (non-)strategic behavior from Wright and Leyton-Brown (2022). In particular, we use their definition of a *weakly non-strategic model*.

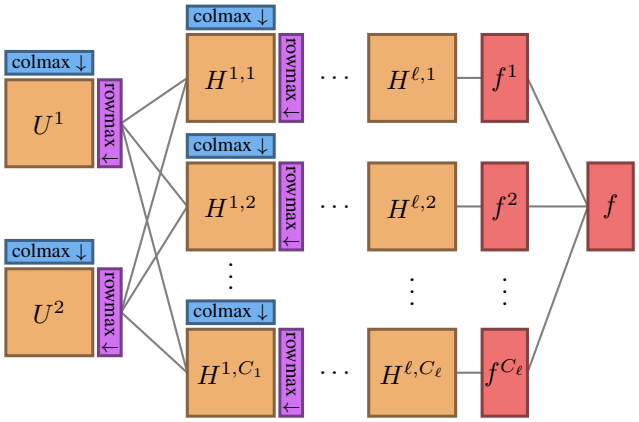

Figure 1: GameNet's feature layers.

Definition 2.3. *An action a_i^+ in a game $G = (A, u)$ is* ζ-dominant *if $u_i(a_i^+, a_{-i}) > u_i(a_i, a_{-i}) + \zeta$ for all $a_i \neq a_i^+$ and $a_{-i} \in A_{-i}$. Then, a behavioral model f_i is* dominance responsive *if there exists some $\zeta > 0$ such that, in all games G with a ζ-dominant action a_i^+, the mode of f_i is a_i^+: that is, $f_i(G)(a_i^+) > f_i(G)(a_i)$ for all $a_i \neq a_i^+$.*

Definition 2.4. *A behavioral model f_i is* weakly non-strategic *if it cannot be represented as quantal best response to some dominance-responsive model f_{-i}.*

Wright and Leyton-Brown (2022) also define a class of non-strategic behavioral models called *elementary models*. These models independently compute a "potential" for each outcome in the game, then predict an action distribution based only on these potentials. Intuitively, as long as the potential function is well-behaved, compressing two utilities into one potential creates an information bottleneck that makes strategic reasoning impossible.

Definition 2.5. *A function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is* dictatorial *iff it is completely determined by one input: that is, either $\varphi(x, y) = \varphi(x, y')$ for all $x, y, y' \in \mathbb{R}$, or $\varphi(x, y) = \varphi(x', y)$ for all $x, x', y \in \mathbb{R}$.*

Definition 2.6. *A function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is* non-encoding *iff, for all $i \in \{1, 2\}$ and $b > 0$, there exist $x, x' \in \mathbb{R}^2$ such that $\varphi(x) = \varphi(x')$ but $|x_i - x'_i| > b$.*

Definition 2.7. *An* elementary behavioral model *is a model of the form $f_i(G) = h_i(\Phi(G))$, where*

1. *Φ maps an $n \times m$ game G to a potential matrix $\Phi(G)$ by applying a potential function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ to each utility vector $(u_1(a), u_2(a))$;*
2. *φ is either dictatorial or non-encoding; and*
3. *h_i is an arbitrary function mapping an $n \times m$ potential matrix to a vector of n probabilities.*

The key fact that we will leverage is that convex combinations of elementary models are weakly non-strategic.

Theorem 2.8. *(Theorem 5, Wright and Leyton-Brown 2022) Let $g_i^1, \ldots, g_i^K$ be elementary behavioral models, and let the weights $w_1, \ldots, w_K \geq 0$ sum to 1. Then, the behavioral model $f_i(G) = \sum_{k=1}^K w_k g_i^k(G)$ is weakly non-strategic.*

3 GameNet's Feature Layers are Strategic

We now have a framework with which we can formally study GameNet's feature layers. Recall that these feature layers output a predicted level-0 behavior; indeed, they are flexible enough to represent many existing level-0 models, such as the maxmax heuristic. However, this flexibility turns out to be a double-edged sword.

Our first main result is that GameNet's level-0 model can represent a particular strategic model—quantal best response to maxmax—to arbitrary precision. We give a constructive proof, providing parameter values for a 3-layer network that approximates this strategic model.

Theorem 3.1. *Let $q_1(G) = QBR_1(maxmax_2(G); 1, G)$. Let $\mathcal{G}$ be the set of games where all utilities are between 0 and C_{max} and all utilities differ by at least C_{gap}, where $C_{max}, C_{gap} > 0$ are arbitrary constants. Then, there exists an instantiation of GameNet's feature layers that coincides with $q_1(G)$ for all $G \in \mathcal{G}$.*

Proof. We first give a series of computations that produce $q_1(G)$ for all $G \in \mathcal{G}$. Let

$$\begin{aligned} M_c &= \text{colmax}(U^2), \\ M_* &= \text{rowmax}(M_c), \\ B &= \text{relu}(M_c/C_{gap} - M_*/C_{gap} + 1), \\ E &= \text{relu}(U^1 + C_{max}B - C_{max}), \text{and} \\ Q &= \text{softmax}([\textstyle\sum_j E_{i,j}]_{i=1}^n). \end{aligned}$$

Here, M_c is a matrix where each column contains the maximum utility that player 2 could realize by playing that action. M_* is a constant matrix containing player 2's maxmax value. Then, because all utilities differ by at least C_{gap}, B is a matrix containing a column of ones for player 2's maxmax action and zeros in all other columns. (This maxmax action is unique because all of the utilities are distinct.) Because all utilities are between 0 and C_{max}, E is a matrix containing player 1's utilities in player 2's maxmax column and zeros elsewhere. Finally, Q is a vector containing player 1's quantal best response to the maxmax action. Therefore, $Q = q_i(G)$.

We now show that GameNet's feature layers can represent these computations. Consider a model with three hidden layers, with two hidden units in the first two layers and one hidden unit in the final layer. Assume that all unspecified weights and biases are set to zero. In the first layer, setting $w^1_{1,2} = 1$ gives $H^{1,1} = U^1$, and setting $w^1_{2,6} = 1$ gives $H^{1,2} = M_c$. In the second layer, setting $w^2_{1,1} = 1$ gives $H^{2,1} = U^1$, and setting $w^2_{2,2} = 1/C_{gap}$, $w^2_{2,4} = -1/C_{gap}$, and $b^2_2 = 1$ gives $H^{2,2} = B$. In the last layer, setting $w^3_{1,1} = 1$, $w^3_{1,2} = C_{max}$, and $b^3_1 = -C_{max}$ gives $H^{3,1} = E$. The softmax operations at the end of the feature layers complete the proof. □

Note that the assumption that the games in $\mathcal{G}$ have positive utilities was made purely for clarity of exposition. It is straightforward to handle negative utilities by adding appropriate constant shifts. We prove this more general claim in the extended version of the paper.

4 A Non-Strategic Neural Network

Having seen that GameNet's feature layers can represent strategic reasoning, we now introduce a new neural network architecture that is only capable of non-strategic behavior.

ElementaryNet

In Theorem 2.8, we saw that it is possible to construct non-strategic behavioral models by composing an arbitrary response function with a potential function, as long as this potential function is either dictatorial or non-encoding. Intuitively, as long as the potential function is nicely behaved, it adds an information bottleneck to the model, discarding information about the agents' utilities before the response function can perform more complex computations.

This inspires ElementaryNet, a new architecture that adds such an information bottleneck to GameNet's flexible feature layers. ElementaryNet is a model of the form $f_i(G) = \sum_{p=1}^{P} w_p \cdot h^p_i(\Phi^p(G))$, where $\Phi^p : \mathbb{R}^2 \to \mathbb{R}$ are parameterized potential functions applied elementwise to the utility matrices; h^p are parameterized response functions analogous to GameNet's feature layers; and w_p is a vector of probabilities. Figure 2 illustrates this architecture.

We will study two particular instantiations of this model, differing in the specification of their potential functions. The first instantiation, which we dub the *learned-potential* model, allows the potentials to be arbitrary linear functions

$$\varphi^p(x, y) = \theta^p_x x + \theta^p_y y,$$

where the coefficients θ^p_x and θ^p_y are trainable parameters. This model is therefore able to use any linear potential function, so long as it is justified by the training data.

The second instantiation, which we dub the *fixed-potential* model, uses four fixed linear potential functions:

$$\begin{aligned} \varphi_{\text{own}}(x, y) &= x, \\ \varphi_{\text{opp}}(x, y) &= y, \\ \varphi_{\text{sum}}(x, y) &= x + y, \text{and} \\ \varphi_{\text{diff}}(x, y) &= x - y. \end{aligned}$$

This set of potential functions is natural. Noticing that the sign of the potentials is unimportant (as the response function can easily negate all potential values before applying any other computation), these are the four distinct linear potential functions that can be constructed using the coefficients -1, 0, and 1. All four also have an economic interpretation. φ_{own} describes "single-agent" reasoning about the agent's own payoffs; φ_{opp} describes purely altruistic reasoning about the opponent's payoffs; φ_{sum} computes the welfare of each outcome; and φ_{diff} measures the (un)fairness of each outcome. Accordingly, these potential functions—specifically, φ_{own}, φ_{sum}, and φ_{diff}—can be used to represent all of the level-0 heuristics from Wright and Leyton-Brown (2019). A fixed-potential ElementaryNet model has no trainable parameters in its potential functions; its parameters consist solely of those in the response functions h^p and the convex combination w_p.

ElementaryNet is Non-Strategic

Our second theoretical result is that ElementaryNet is weakly non-strategic: it cannot represent quantal best response to any dominance-responsive behavioral model.

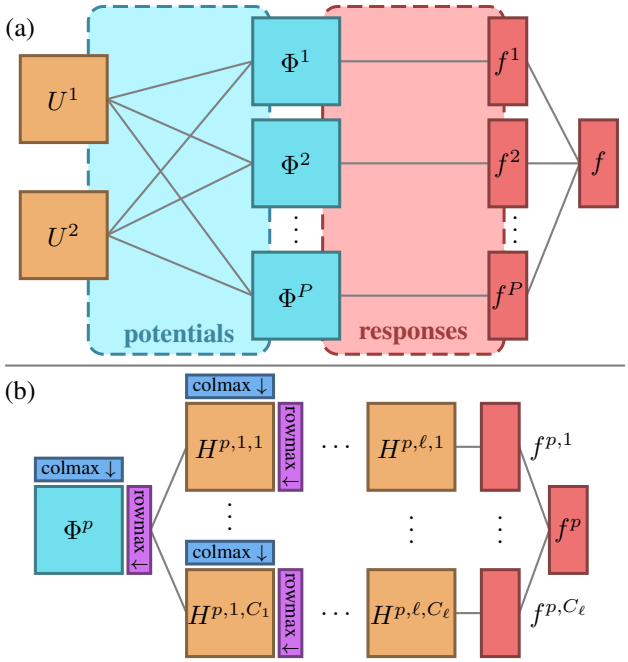

Figure 2: The ElementaryNet architecture. (a) The full model; (b) a representative response function.

Theorem 4.1. *ElementaryNet is weakly non-strategic.*

Proof. We will first show that any linear potential function $\varphi(x, y) = \theta_x x + \theta_y y$ is either dictatorial or non-encoding. If $\theta_x = 0$ or $\theta_y = 0$, then φ is dictatorial. Otherwise, let $b > 0$ be arbitrary. Let $(x_1, y_1) = (0, 0)$ and $(x_2, y_2) = (2b, -2b\frac{\theta_x}{\theta_y})$. Then, we have $\varphi(x_1, y_1) = \varphi(x_2, y_2) = 0$ but $|x_1 - x_2| = 2b > b$. Similarly, let $(x_3, y_3) = (2b\frac{\theta_y}{\theta_x}, -2b)$; we have $\varphi(x_1, y_1) = \varphi(x_3, y_3) = 0$ but $|y_1 - y_3| = 2b > b$. Therefore, φ is non-encoding.

Because each of its potential functions are either dictatorial or non-encoding, ElementaryNet is therefore a convex combination of elementary models. Then, Theorem 2.8 implies that it is weakly non-strategic, completing the proof. □

Intuitively, linear functions have linear level curves, which are either axis-aligned (implying that the function is dictatorial) or extend arbitrarily far in both dimensions (implying that the function is non-encoding). In either case, linear potential functions discard enough information to ensure that the model cannot represent strategic reasoning, regardless of the flexibility of the response functions.

Notably, ElementaryNet does not just disagree with quantal best response on isolated games: it is incapable of representing broad categories of strategic reasoning. It can fail to be "other-responsive", disregarding the other player's preferences entirely; such a model cannot possibly form accurate beliefs about an opponent. Otherwise, if it is other-responsive, it can then be made to play dominated actions with probability bounded away from zero, regardless of how large the losses in utility are. We formalize and prove these properties in the extended version of the paper.

5 Experiments

While the constraints in ElementaryNet's architecture make it provably unable to represent strategic reasoning, only an empirical study can determine the extent to which these constraints affect the model's ability to model human behavior. Here we show the most positive result that we could hope for: that, when ElementaryNet is used as the level-0 specification for an iterative reasoning model, it matches the performance of GameNet. What's more, since our model cleanly factors descriptions of strategic and nonstrategic behavior into different parameters, we can gain insights into human behavior by varying features of the model and analyzing its parameters. We demonstrate how this type of analysis can yield interpretable results, showing that iterative reasoning is a good model of human behavior; that the precise model of iterative reasoning is relatively unimportant; that far simpler level-0 specifications are inconsistent with non-strategic behavior in our data; and that level-0 specifications from prior work lose relatively little predictive power.

Experimental Setup

Data. We used a dataset consisting of results from twelve experimental studies. The first ten were used in past work comprehensively evaluating behavioral game theory models (Wright and Leyton-Brown 2017, 2019). The remaining two studies ran large-scale experiments on Amazon Mechanical Turk (Fudenberg and Liang 2019; Chui, Hartline, and Wright 2023). In total, the dataset contains 26,553 observations across 366 distinct games. We provide more details about this data in the extended version of the paper.

Training. To train our models, we first randomly selected 20% of the games to use as a validation set and 20% to use as a test set, using the remaining 60% as a training set. Then, we trained up to 36 models with different hyperparameters, varying the L1 regularization coefficient applied to the neural network weights, the dropout probability, and the initial QCH model parameters. The exact hyperparameter values we used are detailed in the extended version of the paper. We report the test loss of the model that had the lowest validation loss. We used the squared L2 error between the predicted distribution and empirical distribution as our loss function, as past work has argued that it is appropriate for evaluating behavioral models (d'Eon et al. 2024). We repeated this procedure for 50 train/validation/test splits.

Confidence intervals. Due to the small size of our dataset and the comparative flexibility of our models, the losses depend heavily on precisely which games are in the training, validation, and test sets, adding substantial variance to our training procedure. To combat this high variance, rather than reporting losses of our models, we report the difference in loss to a reference model on the same data split. Taking paired differences in this way removes any variance caused purely by differences in the data, isolating differences in model performance. We report bias-corrected accelerated bootstrapped confidence intervals (Efron 1987) of the differences in test loss with confidence levels of both 68% and 95%. We also report absolute model losses in the extended version.

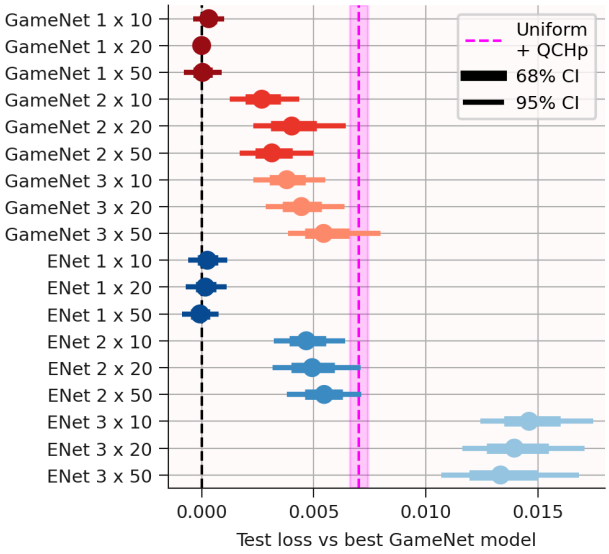

Figure 3: ElementaryNet and GameNet level-0 models trained with a QCH-Poisson strategic model. The best ElementaryNet models are similar in performance to the best GameNet model. Lower values are better.

Comparing to GameNet

First, to evaluate whether ElementaryNet's restrictions harmed its predictive performance, we compared it to GameNet. We trained QCHp models using a variety of GameNet and ElementaryNet models as their level-0 input. The ElementaryNet models each used a single learned potential function. We varied GameNet's feature layers and ElementaryNet's response functions to use 1, 2, or 3 hidden layers with widths of 10, 20, or 50. We also evaluated the Uniform + QCHp model as a baseline.

Our results are shown in Figure 3. Consistent with the results reported by Hartford, Wright, and Leyton-Brown (2016), we found that the best GameNet models performed far better than the Uniform + QCHp baseline. GameNet models with only a single hidden layer performed best, while deeper models had worse test loss, a likely sign of overfitting. However, the best ElementaryNet models—which also had one hidden layer—had nearly identical performance, with test losses that are statistically indistinguishable from the best GameNet model. These results show that, despite the additional constraints on ElementaryNet, it constitutes a state-of-the-art model of human strategic behavior when used in tandem with a QCHp model.

Leveraging our Interpretable Model

Having showed that ElementaryNet + QCHp is a state-of-the-art model of strategic behavior, we then sought to leverage its theoretical guarantees to understand what types of reasoning might be on display in our data. We ran four follow-up experiments varying different features of the model. In each case, we found insights about the kinds of strategic or non-strategic reasoning exhibited by the participants in our dataset.

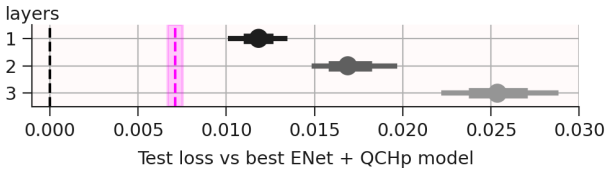

Figure 4: ElementaryNet models with no QCH model. Purely non-strategic models fit the data extremely poorly.

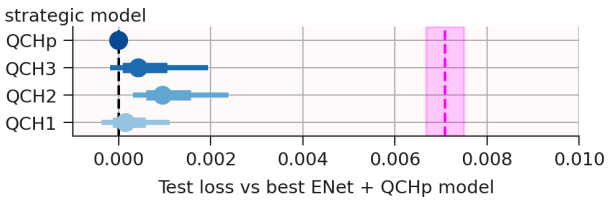

Figure 5: ElementaryNet level-0 models with various QCH models. Changing the structure of the level distribution has little effect on model performance.

Models with no strategic reasoning. We first compared the performance of our best ElementaryNet + QCHp model to one with no QCH model at all. This experiment is similar to the one that Hartford, Wright, and Leyton-Brown (2016) ran, where they found that GameNet performed better with only a level-0 model. However, we showed that GameNet's level-0 model is able to emulate strategic models, making it impossible to conclude that iterative strategic reasoning is a poor model of human strategic behavior from this empirical result. In contrast, we can be sure that ElementaryNet's level-0 model alone is non-strategic.

We trained learned-potential ElementaryNet models with 1, 2, and 3 hidden layers, with 50 hidden units in each layer, and no strategic model. The results (Figure 4) show that these purely non-strategic models fit the data extremely poorly, with test losses significantly worse than even the baseline model. Thus, we can conclude that iterative reasoning is indeed a good model of the behavior of our subjects, far outstripping any non-strategic model.

Different levels of strategic reasoning. Next, we varied the strategic model. We focused on our best performing level-0 model: ElementaryNet with a learned potential and a response function consisting of 1 hidden layer of 50 units. However, we replaced the strategic model with QCH models having arbitrary level distributions up to 1, 2, or 3 levels, allowing these level distributions to be learned during training. The results (Figure 5) show that the differences between the various QCH models were relatively small. ElementaryNet models trained with QCH1 and QCH3 had performance that was statistically indistinguishable from the results with the original model using QCHp. The model using QCH2 performed the worst, but still had an average test loss far better than that of the baseline Uniform + QCHp model.

Each of these QCH models have interpretable level distributions, which are either discrete histograms (for the QCH1, QCH2, and QCH3 models) or Poisson distributions (for the QCHp model). Figure 6 shows the average learned level dis-

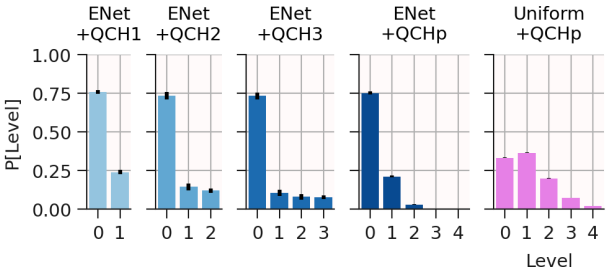

Figure 6: Fitted QCH parameters. Error bars show 95% confidence intervals on parameter values.

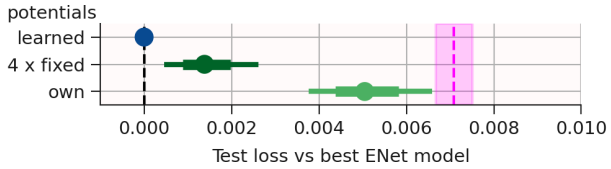

Figure 7: ElementaryNet level-0 models with various potential functions. Overly restrictive and richer specifications of non-strategic behavior both degraded performance.

tribution for each of the ElementaryNet models, as well as the baseline Uniform + QCHp model. These plots show that each of the QCH models co-trained with an ElementaryNet level-0 model place over 70% of their probability on level-0 reasoners, while the baseline model places only 33% of its probability on level-0 reasoners. This difference suggests that the majority of subjects in our data are likely performing some type of relatively rich non-strategic reasoning, which the baseline model is forced to model as imperfect strategic behavior. This qualitative finding echoes previous results from Wright and Leyton-Brown (2019), who found that using level-0 models beyond the uniform baseline increased the fitted proportion of level-0 reasoners from 30% to 60%.

Simpler level-0 specifications. Having studied our strategic model extensively, we next varied features of our level-0 model. A natural question to ask here is whether our nuanced definition of non-strategic behavior was necessary. After all, it would be easier to simply create a level-0 model that makes predictions based only on the agent's own payoffs. Such a model is clearly non-strategic, as it cannot reasonably form beliefs about the opponent's behavior without knowing the opponent's payoffs. Would this simpler level-0 specification be sufficient for modelling our subjects' behavior?

We trained an ElementaryNet model with a single potential function fixed to the "own" potential, which conforms to this more restrictive level-0 specification. As before, we used a response function with one hidden layer of 50 units and a QCHp strategic model. The results (Figure 7, light green) show that this simpler level-0 model performed far worse than our best ElementaryNet model, with average test loss closer to that of the Uniform + QCHp baseline. This provides evidence that the experimental subjects exhibited rich non-strategic behavior that cannot be expressed without knowledge of the opponent's payoffs.

Level-0 potentials from prior work. Lastly, we considered a middle ground: an ElementaryNet model with multiple fixed potential functions. We trained an ElementaryNet model with four potential functions, using each of the four fixed potential functions that we described in Section 4. This model is unable to adapt its potential functions to the data, but can still express richer non-strategic behavior, such as the heuristics used by Wright and Leyton-Brown (2019). As before, the model still used response functions with one hidden layer of 50 units and a QCHp strategic model. The results (Figure 7, dark green) show that the model with four fixed potentials far outperformed the model with just one. However, it still did not match the performance of our best ElementaryNet model, which had just a single learned potential function. This result suggests that it may be possible to develop simple level-0 heuristic models that outperform those from past work by considering more nuanced potential functions beyond welfare and fairness.

6 Conclusions and Discussion

Model performance often comes at the cost of interpretability. In this work, we proved that GameNet, an opaque, state-of-the-art predictor of human strategic behavior, has deep-seated interpretability problems: its purportedly level-0 model is able to emulate strategic reasoning. However, we showed that ElementaryNet, a variation of GameNet with additional restrictions, is completely unable to represent strategic behavior. These restrictions make the model far easier to interpret, making it possible to derive insights about human behavior by inspecting the model's learned parameters or varying its features, without losing predictive power.

We see several promising directions for future work; we describe three here. First, we focused on interpreting the strategic model and the potential functions, two pieces of the model with interpretable parameters, but not the response functions, which are still black boxes. Future work could analyze these response functions by studying how they behave in simple cases, or by testing whether they continue to work well with additional restrictions.

Second, Zhu et al. (2024) presented an alternative method for improving behavioral game theory models, using a neural network to control the parameters of a quantal response model. Fusing their strategic models with ElementaryNet's level-0 predictions could produce interpretable, yet highly predictive, models of human behavior.

Third, we focused on unrepeated, two-player normal-form games. These relatively simple games already evoke strategic behavior, but they do not include other elements such as sequential interactions, random chance, and imperfect information that are common in real-world settings. Extending our models to handle these elements is non-trivial, as it will require developing compelling definitions of non-strategic behavior beyond normal-form games and significantly modifying our architecture. Nonetheless, we are hopeful that the core idea of combining a rich model of non-strategic behavior with a structured model of strategic reasoning will be successful in these more complex settings.

Acknowledgments

This work was funded by an NSERC Discovery Grant, a CIFAR Canada AI Research Chair (Alberta Machine Intelligence Institute), and computational resources provided both by UBC Advanced Research Computing and a Digital Research Alliance of Canada RAC Allocation. Additionally, Greg d'Eon and Hala Murad were supported in part by funding from the UBC Advanced Machine Learning – Training Network, and Hala Murad was supported in part by funding from a UBC Work Learn International Undergraduate Research Award.

References

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3): 861–898.

Chui, D.; Hartline, J.; and Wright, J. R. 2023. Non-strategic econometrics (for initial play). In *AAMAS 2023*, 634–642. ISBN 9781450394321.

Cooper, D. J.; and Van Huyck, J. B. 2003. Evidence on the equivalence of the strategic and extensive form representation of games. *Journal of Economic Theory*, 110(2): 290–308.

Costa-Gomes, M. A.; Crawford, V. P.; and Broseta, B. 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5): 1193–1235.

Costa-Gomes, M. A.; and Weizsäcker, G. 2008. Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3): 729–762.

d'Eon, G.; Greenwood, S.; Leyton-Brown, K.; and Wright, J. R. 2024. How to evaluate behavioral models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397): 171–185.

Fudenberg, D.; and Liang, A. 2019. Predicting and understanding initial play. *American Economic Review*, 109(12): 4112–4141.

Goeree, J. K.; and Holt, C. A. 2001. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5): 1402–1422.

Hartford, J. S.; Wright, J. R.; and Leyton-Brown, K. 2016. Deep learning for predicting human strategic behavior. In *NIPS 2016*.

Haruvy, E.; and Stahl, D. O. 2007. Equilibrium selection and bounded rationality in symmetric normal-form games. *Journal of Economic Behavior & Organization*, 62(1): 98–119.

Haruvy, E.; Stahl, D. O.; and Wilson, P. W. 2001. Modeling and testing for heterogeneity in observed strategic behavior. *Review of Economics and Statistics*, 83(1): 146–157.

Kingma, D. P.; and Ba, J. 2017. Adam: a method for stochastic optimization. arXiv:1412.6980.

Levine, D. K. 1998. Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3): 593–622.

Nagel, R. 1995. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5): 1313–1326.

Rogers, B. W.; Palfrey, T. R.; and Camerer, C. F. 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, 144(4): 1440–1467.

Stahl, D. O.; and Haruvy, E. 2008. Level-n bounded rationality and dominated strategies in normal-form games. *Journal of Economic Behavior & Organization*, 66(2): 226–232.

Stahl, D. O.; and Wilson, P. W. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3): 309–327.

Stahl, D. O.; and Wilson, P. W. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1): 218–254.

Wright, J. R.; and Leyton-Brown, K. 2010. Beyond equilibrium: predicting human behavior in normal-form games. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wright, J. R.; and Leyton-Brown, K. 2017. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior*, 106: 16–37.

Wright, J. R.; and Leyton-Brown, K. 2019. Level-0 models for predicting human behavior in games. *Journal of Artificial Intelligence Research*, 64: 357–383.

Wright, J. R.; and Leyton-Brown, K. 2022. A Formal Separation Between Strategic and Nonstrategic Behavior. arXiv:1812.11571.

Zhu, J.-Q.; Peterson, J. C.; Enke, B.; and Griffiths, T. L. 2024. Capturing the Complexity of Human Strategic Decision-Making with Machine Learning. arXiv:2408.07865.