On the complexity of forming mental models

Chad Kendall
Department of Finance and Business Economics, Marshall School of Business, University of Southern California

Ryan Oprea
Economics Department, University of California

We experimentally study how people form predictive models of simple data generating processes (DGPs), by showing subjects data sets and asking them to predict future outputs. We find that subjects: (i) often fail to predict in this task, indicating a failure to form a model, (ii) often cannot explicitly describe the model they have formed even when successful, and (iii) tend to be attracted to the same, simple models when multiple models fit the data. Examining a number of formal complexity metrics, we find that all three patterns are well organized by metrics suggested by Lipman (1995) and Gabaix (2014) that describe the information processing required to deploy models in prediction.

Keywords. Complexity, mental models, inference, bounded rationality, behavioral economics, economics experiments.

JEL classification. C0, C91, D91.

1. Introduction

In order to understand the world, we continuously have to form *mental models* of data generating processes (DGPs). That is, we have to form beliefs about the causal dependence of outcomes we care about on other observables (e.g., characteristics, past events) and deploy those beliefs in prediction.[1] These inference problems are fundamental, arising regularly across domains in social and economic life. For instance, in order to strategically interact with others, we have to use the history of play to infer the strategy they are using. In order to predict other people's behavior, we have to infer the habits, heuristics, norms, and demand functions that drive their decision-making. In order to make

Chad Kendall: chadkend@marshall.usc.edu
Ryan Oprea: roprea@gmail.com

We thank Guillaume Frechette, Cary Frydman, Thomas Graeber, Ariel Rubinstein, Andrew Schotter, Emanuel Vespa, Georg Weizsacker, and Sevgi Yuksel for valuable conversations about this research. We are grateful to seminar audiences at New York University, Berkeley Haas, Yale SOM, Stanford, Surrey, Queensland, the Berlin Behavioral Economics Seminar, Texas Tech University, the 6th World Congress of the Game Theory Society, the Workshop on Applications of Revealed Preferences (WARP), and USC. This research was supported by the National Science Foundation under Grant SES-1949366 and was approved by UC Santa Barbara IRB.

[1] We refer to "mental models" (rather than "models") to emphasize that they need not be formal, explicit, or even consciously available to decision makers (DMs) in order to guide choice.

Available at http://qeconomics.org. https://doi.org/10.3982/QE2264

political decisions, we have to form models about how policies and geopolitical events conspire to generate the social or economic outcomes we care about. To understand any natural process or solve an engineering problem, we have to form predictive models of nature's laws.

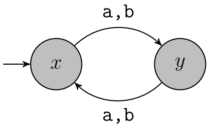
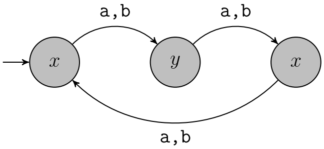
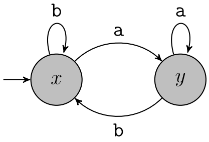
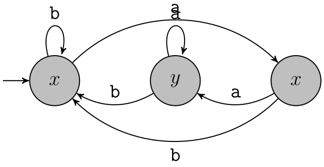
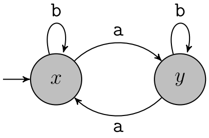
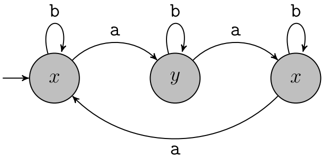
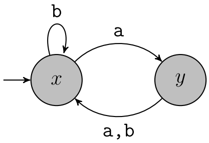
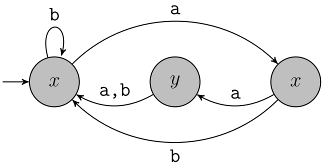
Although there is a long experimental literature studying how and why subjects fail at simple statistical inference (i.e., Bayesian inference problems, e.g., Benjamin (2019)), there has been almost no investigation of the equally fundamental question of how (and how well) people form predictive mental models of the sorts of algorithms that connect variables across social and economic life. This is important because even when data generating processes are perfectly deterministic, requiring no statistical reasoning, they may nonetheless be difficult to accurately model because doing so requires elaborate pattern recognition and sophisticated representation of what has been recognized. We call this difficulty "inferential complexity." What makes a data generating process difficult to infer from data? What makes a coherent model *complex* to construct from the observation of past events? What kinds of models are people good at forming and what kinds of models do people form when many models are available to explain data?

In this paper, we report an experiment designed to take some first steps toward answering these kinds of questions. In order to do this, we reverse the approach typically taken in the inference literature: instead of studying statistical inference in stochastic but algorithmically trivial environments as the literature generally does, we study deterministic but algorithmically complex inference settings. The experiment consists of a series of tasks each consisting of three parts. In part 1, we show subjects a data set—a sequence of twelve *inputs* (a or b repeated in some sequence), and following each, an *output* (e.g., x or y)—generated by a deterministic, but unknown, algorithm that may (or may not) respond to the history of inputs. We tell subjects virtually nothing about the true data generating algorithm other than that it may or not respond to inputs and that it is not random. The subject's job in part 2 is to predict the outputs that this same algorithm will produce after each of twelve further inputs, presented to the subject in sequence. In order to guess effectively, the subject must "extract" a model—a rationalizing algorithm—from the part 1 data (even if unconsciously). Based on their choices in part 2, we can infer whether or not they were successful at forming an accurate model. In part 3, we incentivize subjects to explicitly "articulate," in words, the model they have extracted.

In our main treatment (called "Unique"), we apply this three-stage process to a variety of DGPs, each a distinct, deterministic algorithm connecting inputs to outputs. Specifically, we vary the underlying DGP in a series of eight distinct tasks. Table 1 visualizes these DGPs, representing each as a *finite automaton*. In these automata diagrams, the DGP is visualized as (i) a collection of *states* (a set of circles), (ii) outputs produced by the DGP in each state (shown as letters inside of the circles), and (iii) transitions between states (shown as arcs), triggered by inputs (letters next to these arcs). A free-standing arrow points to the initial state for the DGP.

Even though we constrain ourselves to the simplest algorithms, we are able to confront subjects with a rich variety of data generating processes that arise regularly in social life. For instance, in "autonomous" processes (A2 and A3 in Table 1) outputs cycle independently of the inputs, mirroring, for example, seasonal oscillations in traffic

TABLE 1. Examples of data generating processes (DGPs) described using finite automata.

DGP	2-State DGP	3-State Extension
Autonomous *A fixed pattern, independent of inputs*	 **A2**	 **A3**
Instruct *A direct mapping of previous input to output*	 **I2**	 **I3**
Switch *A direction to move back or forth over a sequence of outputs*	 **S2**	 **S3**
Hybrid *One input Instructs, the other prescribes a Switch*	 **H2**	 **H3**

Note: We include the 4 types of nontrivial (no sink or source states), simple (2-input and 2-output) 2-state automata and examples of a 3-state extension of each of these. These are the DGPs we assign to subjects in the experiment.

patterns or the daily habits of a coworker. By contrast, with "instruction" processes (I2 and I3), the current input directly determines the output, as in the way income predictably affects consumer demand, or the way someone playing a tit-for-tat strategy directly responds to their opponent's past action in a repeated prisoner's dilemma game. In "switch" processes (S2 and S3), the input instead causes a *change* in the output, for example, as in consumers that switch brands when they receive poor customer service, or voters who oust the current incumbent whenever the economy is bad.[2] Our algo-

[2] Of course, our interest is in how successful people are at recognizing and distinguishing between these and other DGPs. To take the last example, the question is how difficult it is for a politician to discover that voters change their vote in response to a bad economy (a "switch" process) when other possibilities exist

rithms include these DGPs and more elaborate DGPs that generate outputs based on more intricate patterns in the history of inputs.

We find strong evidence that extracting a correct model from data and using it to guide prediction is quite difficult for the average subject. Subjects are only able to extract a model that rationalizes the part 1 input string about 40% of the time.[3] What is more is that this "extraction rate" is highly variable across the types of DGPs we study, with some inducing extraction rates as high as 76% and others as low as 5%. Crucially, we find that this variation has structure both within and between subjects: subjects that extract models of on-average difficult DGPs tend to extract on-average easier DGPs, too. This regularity suggests that a latent complexity ordering over DGPs exists, governing failures to extract models across subjects.

The richness of the set of algorithms we study allows us to look for formal structure in this complexity ordering by evaluating the predictive power of a number of notions of complexity culled from various disciplines. We cast a wide net, collecting notions and metrics from computer science, information theory, algorithmic information theory, and economics. Some of these notions are rooted in characteristics of the DGPs themselves and the implementation costs they exact—for instance, the minimum number of states (state complexity) or transitions (transition complexity) in the simplest description of the algorithm, or the minimal resource burden of implementing the algorithm on physical hardware. Others, such as measures of the Kolmogorov complexity or the Shannon entropy of the output or input string, are instead rooted in characteristics of the data sets subjects observe. Still others such as the mutual information between inputs and outputs, the sensitivity of outputs to past inputs, and measures of the fineness with which decision makers must partition ("partition complexity") or (relatedly) the number of inputs/outputs that must be attended to ("sparsity complexity") in order to apply a model, are rooted in both the DGP and the kinds of data sets they produce.

We find that two closely related notions that describe the information processing required to apply a model—"partition complexity" (adapted from Lipman (1995)) and "sparsity complexity" (adapted from Gabaix (2014))—are the best at predicting extraction, with impressively strong correlations in excess of 2/3. By contrast, some other measures (e.g., mutual information, entropy, Kolmogorov complexity) do quite poorly, with correlations with extraction of well below 0.5. Others, such as computational complexity and especially the number of transitions in the simplest description of the DGP as a finite state machine, do somewhat better. Perhaps surprisingly, the number of states in the simplest finite state machine description of the DGP—a commonly used measure of

ex ante. Voters may instead simply tire of incumbents so that the process is "autonomous," or, they may vote for one party when the economy is good and the other when it is bad, following an "instruction" rule. Which is the politician most likely to recognize? When multiple processes are consistent with the data, which model is the politician most likely to adopt?

[3] In the Unique treatment, we say that a subject has extracted a model if she guesses correctly in part 2 of the experiment. Of course, a subject may fail in this task simply because she forms a model that is more complicated than the true one, leading to mistakes. This possibility is one of the key reasons we asked subjects to verbally articulate their model in part 3 of each task. Examining this data, we find no evidence that subjects extract models that rationalize the part 1 data set but that require more states than the true DGP.

the complexity of algorithms in other contexts—does a poor job of organizing the difficulty of forming models.

In addition to studying people's ability to "extract" a model to forecast, we also study people's ability to consciously "articulate" a model. We find strong evidence that subjects are often able to extract models that they are unable to explicitly describe. That is, a lot of the learning involved in model formation is implicit rather than explicit. Nonetheless, we find that the explanatory power of complexity notions is nearly identical for articulation and extraction: our information processing measures (partition and sparsity complexity) are highly correlated with articulation rates across DGPs, just as they are with extraction rates.

In the final step in our investigation, we study what types of models people are drawn to when multiple models are consistent with the data. In our "Multiple" treatment, we ran subjects through an identical series of tasks to those in the Unique treatment, featuring the exact same DGPs. However, unlike in Unique, we deliberately showed subjects part 1 data sets that are consistent with *multiple* distinct DGPs and, therefore, are consistent with multiple distinct mental models.[4] Structurally estimating subjects' models from their part 2 choice behavior, we find evidence that subjects are systematically drawn to *simple* models. Under almost any complexity notion available, subjects tend to "select" the simplest model that rationalizes the data set at a rate higher than random. Of these, once again our information processing notions (partition and sparsity complexity)—the notions which best predict rates of extraction and articulation—also best predict model selection when multiple data-consistent models are available. Conditional on extracting a model, subjects extract the partitions/sparsity-simplest model more than 2/3 of the time (compared to a less than the 1/3 rate predicted for a random model selector).

These findings are relevant to a growing literature in economics on "mental models." Most closely related perhaps is Aragones, Gilboa, Postlewaite, and Schmeidler (2005), which points out that forming a model from a noisy data set ("fact free learning") is computationally complex ("NP Hard") and, therefore, likely to be intractable for decision makers as data sets grow large (as the number of variables increase). We show that model formation can be intractably difficult even under deterministic and minimally-sized (two variable) data sets, and that this is because factors other than the dimensionality of the data set (that can vary even in the simplest data sets) generate severe complexity burdens of their own. Our work is also related to recent theoretical (Esponda and Pouzo (2016), Bohren and Hauser (2021), Fudenberg, Romanyuk, and Strack (2017), Heidhues, Koszegi, and Strack (2018), and Gagnon-Bartsch, Rabin, and Schwarzstein (2021)) and empirical (Hanna, Mullainathan, and Schwartzstein (2014), Handel and Schwartzstein (2018), Esponda, Vespa, and Yuksel (2023), Enke (2020), and Graeber (2023)) work showing that decision makers can get "stuck" in misspecified mental models, and have difficulty revising these models with learning. Our work complements this literature by

[4] Technically, any data set is consistent with an infinite set of mental models. By "unique," we mean unique within the class of models that can be described by finite automata with fewer than four states. Our results from the articulation task show that this refinement is not restrictive—we find no evidence from verbal descriptions that subjects form rationalizing models with more states.

showing that there is predictable structure in the types of models people are likely to get "stuck" in in the first place. Our finding that decision makers are initially "biased" toward models that are simple (in the sense that they require little information processing) may be useful in guiding the construction of models in this literature. Finally, there is a recent theoretical literature studying the implications of people being unable to form accurate causal models (Spiegler (2016), Eliaz and Spiegler (2020)), focusing on DGPs in which the direction of causality must be inferred (i.e., describable by directed acyclic graphs). We study simpler DGPs in which the direction of causality is known, but we emphasize that our methods can (and should) be extended to these more difficult settings.[5]

Our work also relates to a literature in economics studying the effects of complexity on human decision-making. Most relevant is the "automata" literature, which models decision rules and procedures (e.g., strategies in games) as finite automata, simple descriptions of algorithms from computer science (see Lipman (1995) and Chatterjee and Sabourian (2009) for reviews). The main strand of this literature studies the effect of"implementation complexity"—the subjective costs of employing decision rules describable by complex automata—on decision-making (Rubinstein (1986), Abreu and Rubinstein (1988), and Kalai and Stanford (1988)). Oprea (2020) measures these implementation complexity costs experimentally by paying subjects to follow rules describable by automata and then eliciting their willingness pay to avoid being assigned these same rules again in the future. In contrast, our paper builds on a second strand of the automata literature, which studies not the implementation complexity of following automata, but rather the "computational complexity" of *inferring* automata from data (Gilboa (1988), Ben-Porath (1990)). This literature shows theoretically that automata are complex to infer in the sense that they become computationally intractable to detect as the number of states in the underlying automata grows. Our work (i) supports this literature by showing that this inference problem is indeed empirically difficult for decision makers and (ii) extends this literature by showing that it is difficult even for the simplest (2- and 3-state) automata because of sources of complexity unrelated to the size of the automaton.[6,7]

Finally, our work relates to several strands of literature in psychology. First, there is a long literature, beginning with Byrne and Johnson-Laird (1989), examining whether

[5] Related to Eliaz and Spiegler (2020) is Schwartzstein and Sunderam (2021), who study the ability of senders to persuade receivers by supplying them models. Our interest lies instead in how people form models that are not supplied to them by others.

[6] Our paper also relates less directly to work that studies the complexity of other objects than automata. For instance, Murawski and Bossaerts (2016) show that metrics of the computational complexity of instances of optimization problems (knapsack problems) predicts failures to solve them. Abeler and Jäger (2015) show that increasing the complexity of taxes reduces responsiveness to their marginal incentives. Jin, Luca, and Martin (2021) show that people use complexity instrumentally to shroud information.

[7] Our paper also relates to a literature that models bounded rationality by postulating costs associated with metrics of decision problems. For instance, an important literature models the effects of attentional costs, often measured by mutual information, on attention costs (e.g., Sims (2003), Caplin (2016), and Dean and Neligh (2023)). Likewise, Gabaix (2014) assumes that sparser models of decision problems are less costly to use and examines the implication for standard optimization problems in economics. We do not fit or test these models directly but instead examine the predictive power of some of the key underlying metrics they are built around (e.g., mutual information, sparsity).

people form mental models when trying to solve propositional and spatial reasoning problems.[8] By comparison, the DGP-inference problem we study is unambiguously related to mental models (it would be difficult to perform in our task without forming a model of some sort), making it a particularly valuable setting for studying what makes forming such models difficult. Second, our finding that articulation of models is more difficult than extraction of models parallels a large literature in psychology that makes a distinction between "explicit" and "implicit" learning. Most of this literature studies very different settings from ours, uses very different methods, and asks fundamentally different questions. Most closely related to our work is a small literature on "biconditional grammars" (Mathews and Buss (1989) and Shanks, Johnstone, and Staggs (1997)), which attempts to distinguish between implicit and explicit learning.[9] Subjects are given a DGP similar to our "instruct" DGP, and asked to identify a rule that relates one sequence of letters to another ("x" goes with "T," etc.). To identify implicit learning, these papers vary whether or not subjects know such a rule is being applied. Because our primary goal is to understand what makes a DGP difficult to model, we instead vary the DGP, holding fixed subjects' knowledge that there is a DGP. To identify explicit learning, these experiments also ask subjects to articulate the rule, but we do so in an incentive compatible way, arguably leading to more credible data.

The remainder of this paper is organized as follows. In Section 2, we describe the problem we are interested in, the class of algorithms we use to generate models in the experiment, and preview a number of metrics of complexity that form hypotheses for our experiment. In Section 3, we discuss and motivate our experimental design, and in Sections 4 and 5, our empirical results. We conclude in Section 6 with a discussion.

2. Conceptual background and questions

Consider a DM who observes a "data set," D^T, consisting of (i) a variable called "inputs," $u_t \in U$ observed at dates, $t = 1 \dots T$ and (ii) a second variable called "outputs," $v_t \in V$ observed at dates, $t = 0 \dots T$. The data set is formed by a "data generating process" (DGP) that produces an initial output, v_0, and subsequent outputs, v_t, at each date t based on the sequence of inputs prior to t, $u_1 \dots u_t$.

We deliberately focus on what is arguably the simplest class of DGPs: DGPs in which (i) inputs and outputs are discrete, (ii) the mapping between inputs and outputs are deterministic, and (iii) the direction of causality between inputs and outputs runs weakly in a single direction. As we discuss below, we focus on this class of DGPs in order to limit

[8] In propositional reasoning, subjects are faced with a series of logical propositions and attempt to derive the logical conclusion. Spatial reasoning problems are similar except that the statements describe locations of objects (e.g., "C" is in front of "x").

[9] The two most popular tasks are artificial grammar tasks (Reber (1967)) and serial reaction time tasks. In the former, subjects observe test "sentences" constructed from a Markovian process and then attempt to determine whether or not subsequent sentences are consistent with the grammar of not. In the latter, they respond to a visual stimulus as fast as possible (see Jimenez, Vaquero, and Lupianez (2006) for a recent contribution). In both tasks, subjects are not told beforehand that a set of rules is generating the patterns they see so that any ability to do better than chance in subsequent guesses is thought to occur through implicit learning.

the number of forces influencing the DM's inference problem—a choice which allows us to draw sharper conclusions on the drivers of the complexity of inference in a particularly simple domain. Nonetheless, our questions and methods can and should be extended to richer DGPs that include stochasticity (e.g., by studying Markov chains) or unclear causality (by studying directed acyclic graphs) in future work.

DGPs in this class are describable as finite state machines (or finite automata)—among the simplest languages for modeling and describing algorithms in computer science. Although in what follows, we will use automaton descriptions to describe the DGPs we study, we emphasize that these DGPs could equally have been described by a great many alternative formalizations of algorithms (some quite rich and sophisticated). We use automata as a descriptive language here largely for expositional purposes and to help to taxonomize and organize the DGPs we select for our design. Later (in Section 2.3), we will compare how well features of automata descriptions (like states and transitions) compare to a number of alternative formal ways of describing these same DGPs in predicting complexity responses.

Formally, an automaton, M, is a four-tuple, (S, s^0, f, τ), where S is a finite set of states, $s^0 \in S$ is the initial state, $f : s \to V$ is an output function that specifies the output $v \in V$ in each state, and $\tau : S \times U \to S$ is a transition function that maps the current state and input, $u \in U$, into the subsequent state.[10] Table 1 contains a sample of eight DGPs, diagrammed using automata descriptions. In these diagrams, (i) each circle represents a state, $s \in S$, (ii) each arc a transition, (iii) letters next to arcs show inputs $u \in U$, and (iv) letters inside circles represent outputs, $v \in V$, in each state. A freestanding arrow pointing to one of the states indicates the initial state, s^0. For instance, automaton I2 (the 2-state "Instruct" DGP) specifies that (i) output x appears at date 0 and (ii) afterwards output x always appears in the period following input b while y always follows input a.

2.1 *Examples*

Anticipating our experiment, the examples in Table 1 are restricted to DGPs that are nontrivial for our purposes—ones describable by "fully connected" automata with no source states (states which are entered but never exited) and no sink states (those which, once entered, are never left).[11] Likewise, we restrict attention to DGPs with the *simplest* input and output alphabets, each containing only two values (e.g., $U = \{a, b\}$ and $V = \{x, y\}$).

The left column in Table 1 shows the four canonical DGPs that are describable by 2-state automata and have only two inputs and outputs (modulo swaps of labels). A2 describes an "autonomous" process that is unresponsive to inputs, switching back and forth between x and y. I2 describes an *instruct* process: each input value uniquely determines the subsequent output value, creating a direct mapping (an instruction) between

[10] Technically, an automaton with such a description is a Moore machine whose outputs only depend on the state. Mealy machines allow outputs to also depend on the current input. There is no loss of generality in focusing on Moore machines as every Mealy machine also has a Moore machine description.

[11] Formally, any automaton can be described as a directed edge graph. We restrict attention to automata for which the associated directed graph is fully connected: every state is reachable from every other.

the previous input and current output. S2 describes a *switch* process: input b leaves the output unchanged, while input a toggles the output, forcing the output to change from its previous value. Finally, H2 describes a *hybrid* process in which one input (b) performs an instruct function while the other (a) performs a switch function.

These four DGPs summarize common processes that arise in inference problems. Autonomous process (A2) are patterns in outputs that are causally distinct from any other variables—simple regularities like the change from day to night or cyclical fads in consumer tastes. Instruct processes (I2) are simple causal relationships that reliably dictate the conditions under which outcomes occur—like the predictable way a person responds to incentives or a chemical is required to generate a reaction. Switch processes (S2) are processes that do not directly cause outcomes, but instead cause changes in outcomes—like the way investors may flip between value and growth stocks when past performance is bad or foraging animals may switch food sources whenever food is scarce. Hybrid (H2) processes blend instruct and switch processes.

In the right-hand column of Table 1, we show the most direct extensions of these four basic processes to richer DGPs—richer in the sense that they require 3 states to describe with automata instead of 2 states. These rules maintain the nontriviality (no sinks or sources) and simplicity (a 2-value input and output library) but require an additional state grafted to each of the four basic 2-state processes.[12,13]

2.2 *Forming models*

Our interest is in understanding how a DM who observes a data set, $D^T \equiv \{u_t\}_{t=1}^T \cup \{v_t\}_{t=0}^T$, generated by an exogenous input string, $\{u_t\}_{t=1}^T$, and DGP, M, forms a "model" of the DGP. We mean three distinct things by "form a model":

Extraction: Suppose a DM after observing D^T observes a further sequence of inputs $\{u_t\}_{t=1+T}^{2T}$ and must guess or forecast the subsequent output sequence $\{v_t\}_{t=1+T}^{2T}$. We say that a DM has "extracted" a model, M, of the DGP if she can correctly list the further output sequence that M produces in response to the continuation of the input string. This ability to reliably "imitate" a DGP is evidence of implicit learning or implicit formation of a model.

Articulation: Suppose a DM after observing D^T is asked to directly describe the DGP M (e.g., using words). We say that a DM has "articulated" a data-consistent model of the DGP if she can correctly describe M. In order to do this, the DM has to be conscious of M. Our distinction between "articulation" and "extraction" is not natural in economic theory, but it mirrors a long-made distinction between "explicit learning" (learning of something one is aware of) and "implicit learning" in psychology (Reber (1967)).

Selection: Suppose a DM observes a D^T, which can be rationalized by any of a set of models/DGPs, $M \in \mathcal{M}$, in the sense that the output sequence $\{v_t\}_{t=0}^T$ could be generated

[12] An alternative way of extending our four 2-state DGPs to 3 states would be to add another output value ("z"). Anticipating our experiment, we opted against this to avoid simultaneously changing two things when moving from 2-state to 3-state DGPs (the size of the output alphabet and the number of states).

[13] Note that unlike DGPs describable as 2-state automata, which have only four basic types (among nontrivial, 2 input/output DGPs), 3-state DGPs have 356. We therefore selected 3-state algorithms that seem most similar to the four 2-state exemplars in the left column of Table 1.

by the input sequence $\{u_t\}_{t=1}^T$, $\forall M \in \mathcal{M}$. We say that the DM has "selected" a model if she has successfully extracted or articulated an $M \in \mathcal{M}$ (even if it is not the "true" model of the DGP ex post). Here, too, the DM has "formed a model" but has had to choose between several candidates. One of our interests in the experiment described below is to understand how a DM chooses between or searches over the multiple rationalizing models in $\mathcal{M}$.

Our aim is to study and compare each of these three aspects of model formation.

2.3 *Complexity*

We are interested in understanding (i) what makes it more difficult to form a model of one DGP than another (i.e., what drives the rates of "extraction" and "articulation" across DGPs) and (ii) what makes one model more likely to be formed when more than one model is consistent with the data set (i.e., what drives "selection").

We are particularly interested in the possibility that either or both might be driven by the "complexity" of the inference task itself. By this, we mean that there is an ordering across DGPs $\succsim_c$ such that if $M' \succsim_c M$, M is easier for any DM to extract or articulate than M', and there is a (possibly different) ordering, $\succsim_{c'}$, such that M is more likely to be selected by any DM than M' when $M, M' \in \mathcal{M}$.

The existence of a consistent complexity ordering is hardly obvious—many other possibilities exist. For instance, it may be that individual DMs instead have idiosyncratic and heterogeneous talents or proclivities for extracting/articulating models of some types of DGPs over others. We call this alternative possibility "affinity" (in the sense that individual DMs may have different individual affinities for identifying some DGPs over others). Even if, in the aggregate, rates of extraction or articulation differ across DGPs, this may be driven not by a consistent complexity ordering over DGPs, but instead by different affinities being more or less common in the population.

Likewise, it is not obvious, ex ante, that selection should be governed by a systematic preferential ordering over models. While it is possible that DMs are systematically drawn to some types of models over others, it is also possible that something more closely resembling random search governs selection. In this case, what matters for selection might be random, with DMs being more likely to successfully infer models whenever more models are available. And even if there is a systematic preferential ordering over models governing selection, it is possible that it is a different ordering from the complexity ordering driving the difficulty of extraction and selection.

Finally, if there is such a thing as a consistent complexity ordering governing model formation, we are interested in understanding what generates that complexity and predicts it. Is there a metric on DGPs (or data sets) that organizes the difficulty of forming models? Does this same metric describe how people choose between equally justifiable models when more than one is available? There are many ex ante notions of complexity (some related to automata descriptions and some completely unrelated) we might consider, and in Section 5.1 we consider a number of them, drawing notions from information theory, computer science, and economics.

3. Experimental design

3.1 *The experimental task*

Each session of the experiment consists of eight *tasks*. Each task consists of three parts, described below.

3.1.1 *Part 1: Observing the data set* In part 1 of each task, we show subjects twelve *inputs* (letters a or b) one at a time on their screens. The inputs are drawn independently with equal probability of each letter. After each input appears, a downward arrow and an output (a letter x or y) appears below it. The outputs are determined by one of the DGPs from Table 1 (varied across tasks) based on the history of inputs. We inform the subject that the input sequence is entirely exogenous, but that outputs are determined by a rule (possibly linked to the prior input sequence), implemented by the computer. Subjects are told nothing about the rule except that (i) it does not choose outputs randomly and (ii) outputs may or may not depend on the preceding history of inputs under the rule. We specify (ii) to avoid deception since two of our rules (A2 and A3) do not depend on prior inputs, and we did not tell subjects that outputs can depend on past outputs because it is technically incorrect—outputs depend on the state, which does not map one-to-one to an output in the 3-state rules.

We call the set of input/output pairs for part 1, the task's "data set."

3.1.2 *Part 2: Extracting a model of the DGP* In part 2 of each task, we show subjects a 13th input and ask them to guess which output comes next in the sequence. They submit their guess by typing a letter (x, y), which shows up on the screen after it is typed. Immediately after typing her guess, a 14th input appears and the subject must guess the next output in the sequence. Inputs continue to appear on the subject's screen until she has observed 12 inputs in part 2 (24 inputs in total including those from part 1) and submitted 12 guesses.

If the task is selected for payment, the subject is paid \$0.35 for every output she guesses correctly. Importantly, the subject *is not shown* the correct sequence (or given any feedback on the correctness of her guesses) until part 3 of the task has concluded, long after she has made all of her part 2 guesses.

We call the subject's ability to repeat the part 1 DGP in part 2, "extraction," and we say that a subject has successfully extracted a model if she chooses a sequence of guesses that is consistent with a DGP that (i) can rationalize the part 1 data set and (ii) is describable by the class of automata with less than four states and no source or sink states. (In Section 4.4, we show that criterion (ii) is not binding in our data: subjects essentially never describe models that satisfy (i) but violate (ii).)

3.1.3 *Part 3: Articulating a model of the DGP* In part 3 of each task, we induce the subject to describe ("articulate") the model she has formed (if any) to explain the data set of inputs and outputs in part 1, by asking her to describe the rule she thinks the computer used to generate the part 1 sequence of outputs. The subject types a description of the rule in a text box on her screen.

The subject is told that this description will (with 10% likelihood) be given to a future participant who will face a guessing task similar to part 2, with outputs determined by the same rule but in response to a different sequence of inputs. That future participant will not have access to a part 1 data set and will therefore rely exclusively on the subject's description of the rule to guide her guesses. The future participant therefore will only be able to correctly guess the sequence of outputs if the subject correctly and clearly articulates a model of the data generating process.

To incentivize the subject to articulate the DGP/rule she inferred from part 1, we pay her $0.35 for each output the future participant manages to guess correctly after reading the subject's description of the rule.[14] In this, we follow previous experimental work incentivizing advice (e.g., Schotter and Sopher (2003)).

3.2 *Tasks and treatments*

We collected data in two treatments that differ only in the specific data sets we assigned to subjects.

3.2.1 *The unique treatment* In the first treatment—called "Unique"—we assign each subject eight tasks, each consisting of a data set (input/output sequence) that can only be rationalized by one, unique, DGP that is describable by a fully connected automaton of fewer than four states. We have two aims with this treatment. The first is to study the degree to which subjects can extract and articulate the unique model that describes the data generating process and how these rates of extraction and articulation vary with characteristics of the generating rule. The second is to study the degree to which subjects' ability to extract and articulate a model varies with characteristics of the data set presented to them.

In Unique, we studied the eight data generating processes listed in Table 1 and discussed in Section 2.1. For each DGP, we created four distinct data sets and assigned each to 1/4 of subjects. Thus, in the Unique treatment we vary DGPs within subject and for each DGP vary data sets between subject.

3.2.2 *The multiple treatment* In the second treatment—called "Multiple"—we inverted the design of the Unique treatment. We assigned subjects the same eight DGPs as in Unique, but gave all subjects part 1 input/output sequences that could be rationalized by *multiple* 2- and 3-state models/DGPs. In Multiple, therefore, subjects could extract any one of a number of simple models, each of which is consistent with part 1 evidence (but only one of which is ex post correct in part 2).[15] The number of such models capable of rationalizing the data set ranged between 2 and 20 depending on the task. Our goal with the Multiple treatment is to study how subjects choose among models when multiple models can rationalize the data set—a model-formation task we call "selection." To collect enough data to credibly answer this question, we gave all subjects in Multiple the same data set for each DGP.

[14] We took pains to assure subjects that the future participant would be sophisticated enough to understand a well-articulated rule in order to further incentivize effort. Specifically, we told them that the participant would be a graduate student in a quantitative field. In practice, we gave each description selected for payment to a single graduate student whose guesses were then used to determine payment.

[15] We provide an example of a sequence that is consistent with multiple models in Supplemental Appendix C (Kendall and Oprea (2024)).

3.3 *Implementation*

We ran the experiment in June of 2021 on the online research panel Prolific using custom Javascript programmed by the authors and deployed using Qualtrics.[16] Each session consisted of detailed instructions (reproduced in Supplemental Appendix D) explaining the task, followed by a set of comprehension questions that subjects were required to correctly answer before moving on to the experiment. Subjects experienced the eight DGPs in a random order across eight tasks. In the Unique treatment, each subject was, with equal likelihood, assigned one of four distinct data sets in each DGP. Experimental sessions took on average 31 minutes. Subjects on average earned $5.63 including the showup fee, resulting in an hourly rate of $10.90, which greatly exceeded Prolific's minimum required rate at the time ($6.50). We targeted 100 subjects per treatment and in the end collected data for 99 subjects in Unique and 101 subjects in Multiple.

A natural concern that applies to any data set collected online is whether some of our results are driven by a lack of attention by online subjects. Several strands of evidence confirm that subjects were attentive both to the instructions and throughout the experiment. First, nearly 70% of subjects made zero errors in our instructions comprehension questions, and 85% made no more than one error. Second, 94% of subjects successfully extracted at least one DGP in part 2 of the experiment—a feat that would be virtually impossible if subjects were not paying attention or misunderstood the instructions. Third, 84% of subjects successfully described at least one DGP in words in part 3 of the experiment—a feat that would likewise be seemingly impossible for inattentive subjects.

3.4 *Understanding the design*

We designed this experiment with several inferential goals in mind.

First, we wanted to directly measure how difficult it is to form models of data generating processes and we designed the Unique treatment with this measurement in mind. By providing subjects with a setting in which there is only one simple model to explain the data, we can directly study the mapping between the true data generating process and the difficulty of forming a model of it.

Second, we wanted to separately evaluate the difficulty of inferring a model implicitly (to guide future behavior) and explicitly (to formally describe it). For this reason, we included both part 2 (in which subjects only have to reproduce and imitate the pattern of the data generating process) and part 3 (in which subjects have to be conscious enough of their model of the DGP to articulate it in words) in each task in the design. Including both allows us to directly contrast what we call "extraction" of a model with "articulation" of the same model, within subject. Furthermore, because subjects are free to describe any model they like in part 3, we can see whether or not it is restrictive to test for extraction within the class of models describable by automata with less than four states and no sink or source states.

[16] We restricted participation to residents of the U.S. and those that had not previously participated (only). The average age of participants was 31.7 and 46.6% were male.

Third, we wanted to study what formalizable features of DGPs—what notions of complexity from literatures in computer science, information theory, algorithmic information theory, and economics—make them more and less difficult (“complex”) for humans to model. To achieve this, we attempted to vary DGPs in a way that plausibly varied difficulty, given recent work in the literature. We achieved this by (i) deploying DGPs that are describable by the canonical four types of 2-state automata and (ii) the most obvious 3-state describable extensions of these four canonical types. This allowed us to vary states in the automaton description (an important driver of implementation complexity in past work; see Oprea (2020)) and also generated significant variation in a number of other complexity metrics that are unrelated to automaton descriptions.

Fourth, although we wanted to vary difficulty, we deliberately limited ourselves to the simplest class of model formation problems available on several margins. For instance, in order to focus on the complexity of extracting patterns, we studied purely deterministic problems in which explicit statistical reasoning is not required. Our methods can easily be extended to study how model formation and its complexity is impacted by stochasticity by, for example, applying them to more general Markov chains. Likewise, we simplified our problems significantly by making the direction of causality clear. But our methods can be easily adapted to study Directed Acyclic Graphs (DAGs), which are typically used to describe causal relationships. We limited ourselves to DGPs with two possible inputs and two possible outputs. But, again, our exact methods can be extended to richer data sets.

Fifth, we wanted to ensure that our results were not overfitted to idiosyncracies in the data sets we assigned to subjects in part 1. Varying the data set also allowed us to study the performance of complexity notions that depend on the details of the data set rather than the underlying DGP. For this purpose, we need sufficient data for each data set. Striking a balance between these two goals, we generated four data sets for each DGP in the Unique treatment and assigned them evenly across subjects.

Sixth, we wanted to study not only what makes a given DGP difficult to model but also what models people are drawn to when more than one can rationalize the data. In particular, we wanted to study the degree to which these two model-formation tasks—extraction and selection—are related to one another. To accomplish this, we introduced a second treatment (“Multiple”) with the same true DGPs as the Unique treatment, but with part 1 data sets that could be rationalized with more than one simple model/DPG.

Finally, we parameterized our tasks so that the part 1 data set and the part 2 guessing task each included 12 inputs with corresponding outputs. Twelve inputs/outputs in the part 1 data set struck a balance between being long enough to be able to find data sets that ensure unique models for the Unique treatment, but not so long as to preclude data sets consistent with multiple models for the Multiple treatment. Twelve inputs/outputs in the part 2 data set is roughly the smallest number we need to structurally distinguish between models that are consistent with the part 1 data set in the Multiple treatment, and was chosen over longer data sets to avoid decision fatigue among subjects. The input sequences were chosen randomly, subject to two constraints: (i) in part 1, they were consistent with only one model (Unique) or more than one model (Multiple) and (ii) in part 2, they uniquely identified the model (Unique and Multiple).

4. Results

4.1 *Extraction*

We first examine the rate at which subjects successfully "extract" a model from the data set they observe. We say that a subject "extracts a model" if she makes choices in part 2 that are consistent with a DGP that can rationalize the data set provided in part 1. To operationalize this, we say that a DGP (and by extension a model) is "data-consistent" if it generates the outputs in the part 1 data set, given the sequence of inputs in that data set. We say that a subject has successfully extracted a "data-consistent model" if a binomial test can reject, at the 1% level,[17] the hypothesis that the subject made choices in part 2 that are random relative to a data-consistent model with fewer than four states.[18,19]

We will say a DGP is relatively "complex" if extracting a model of it proves difficult, on average (though we will revisit the appropriateness of this terminology in Section 4.2).

We focus first on the Unique treatment. Figure 1 plots the extraction rate from the Unique treatment in dark gray for each task (indexed by the true DGP for that task). Small dark dots show the extraction rates for each of the four data sets assigned across subjects in the Unique treatment (in order to visualize variation in extraction rates across data sets for each DGP). Table 2 gives a complementary overview of the data.

The results show that overall subjects successfully extract a model less than half (41%) of the time. However, this global mean conceals substantial heterogeneity across DGPs. Extraction rates vary from very high (over 75% for autonomous processes) to almost zero (as low as 5% for switch processes). This heterogeneity is highly significant—we can reject the hypothesis that extraction rates are identical across DGPs at the 0.0001 level (chi-squared test).[20,21]

Result 1. *Subjects successfully extract models* 41% *of the time, and this rate varies across DGPs between* 76% (*for the "simplest" DGP*) *and* 5% (*for the most "complex"*).

We defer detailed discussion of what characteristics of DGPs and data sets drive this heterogeneity across tasks, but note here one striking fact. The number of states in the DGP (which prior work, Oprea (2020), suggests has a first-order impact on the costs of *implementing* a rule) seems to have only modest influence over the difficulty of *inferring* a DGP. Subjects extract 3-state DGPs 17 percentage points less often than 2-state DGPs (32% vs. 49%). But this is small compared to the variation in extraction rates within state. Indeed, the difference between extraction rates for the most- and least-complex models within state is 60 percentage points, four times larger than the rate difference between

[17] In order to pass this threshold for extraction, the subject had to correctly forecast 11 out of the 12 part 2 outputs.

[18] We show in Section 4.4 that this restriction is not binding in our data.

[19] In forming and interpreting this test, the natural upper bound is 100% data-consistent choices because we deliberately selected data sets in which at least one DGP is fully revealed by the input/output pairs.

[20] In Supplemental Appendix A.3, we replicate this result with a regression analysis that includes DGP, task order, and subject fixed effects.

[21] There is, by contrast, no evidence of order effects in extraction rates. The correlation between task number and extraction rate in the Unique data is -0.026 ($p = 0.459$).

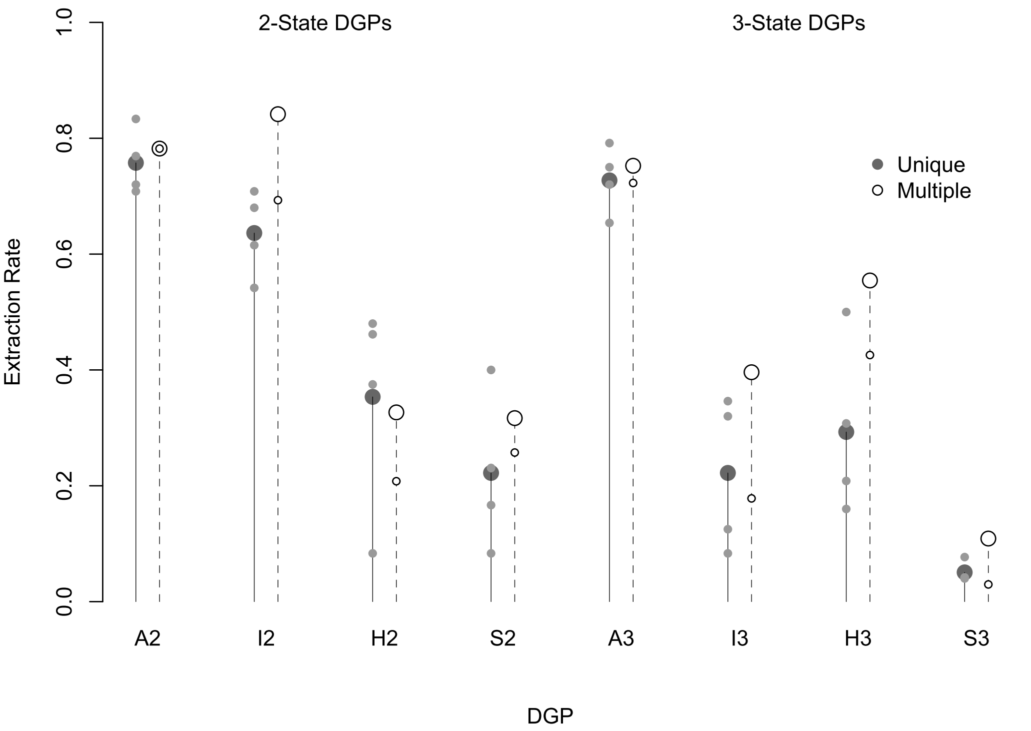


FIGURE 1. Mean extraction rates by DGP and treatment. *Notes*: Data is from 1600 choices in the Unique and Multiple treatments. Large dark dots give the rate at which subjects extracted a model in the Unique treatment. Small dark dots break this rate down by data set. Large hollow dots give the rate at which subjects extracted *any* data-consistent model in the Multiple treatment. Small hollow dots give the rate in which subjects in the Multiple treatment extracted a model of the (ex post) correct DGP.

states. Clearly, complexity of inference is, in large part, driven by factors other than state counts.

Next, we look at how variation across data sets influences extraction rates. Recall, in the Unique treatment that we assigned four different data sets across subjects, each generated by the same DGP but with different input sequences. This variation allows us to ask whether variation in the data set itself generates variation in complexity, holding the DGP constant. The small gray dots in Figure 1 visualize this variation (for each generating model there are four distinct data sets assigned to subjects, producing four small dots).

Under many DGPs, we see little variation in extraction across data sets. For instance, extraction rates are similar (small gray dots are close together) across data sets under the autonomous DGPs (A2 and A3), the 2-state instruct DGP (I2) and the 3-state switch DGP (S3). To formally evaluate this, we ran chi-squared models on each DGP-type separately. Doing so, we can reject the null hypothesis that extraction rates are invariant to

Table 2. Summary statistics.

	Unique			Multiple			
DGP	Extraction	Time	Articulation	Extraction	Time	Articulation	Extraction (True)
A2	0.760	39	0.690	0.780	38	0.730	0.780
A3	0.730	60	0.560	0.750	65	0.520	0.720
H2	0.350	61	0.200	0.330	68	0.130	0.210
H3	0.290	68	0.080	0.550	69	0.160	0.430
I2	0.640	47	0.560	0.840	37	0.660	0.690
I3	0.220	63	0.150	0.400	64	0.250	0.180
S2	0.220	72	0.190	0.320	94	0.170	0.260
S3	0.050	103	0.020	0.110	129	0.040	0.030

Note: Data is from 3200 choices in the Unique and Multiple treatments. Statistics include the extraction rate, the decision time in part 2 in the experiment and the articulation rate for both the Unique and Multiple treatments. For Multiple, we also include the rate at which the subject extracted the true DGP.

the data set at the 5% level only for the hybrid models (H2 and H3); we can reject the same hypothesis at only the 10% level for I3 and S2.[22,23]

Thus, although there is some suggestive evidence that characteristics of the data set may sometimes influence the complexity of extraction, this effect is very small compared to the influence of the underlying DGP itself. Overall, the mean difference in extraction rates between the least and most complex data sets is 22 percentage points, which is less than a third the size of the mean difference between rates for the least and most complex DGPs (70 percentage points). We conclude the following.

Result 2. *Variation in data sets has a much smaller influence over the complexity of extraction than variation in the underlying DGP.*

In Section 5, we consider the relative influence of DGPs and data sets on model formation in more depth, comparing formal measures of complexity that depend on the data set and measures that depend only on the DGP.

4.2 *Complexity*

So far, we have interpreted a low extraction rate as an indication that a DGP is itself "complex" (e.g., difficult or costly). The assumption underlying this interpretation is (i) that there is a latent cost or difficulty ordering across the DGPs themselves and that (ii) subjects differ in their willingness to bear cognitive costs or in their ability to solve problems so that (iii) extraction rates differ on average across tasks.

But, as we discuss in Section 2.3, this "complexity interpretation" may be wrong. For instance, an appealing alternative possibility is that (i) individual subjects have idiosyncratic affinities for inferring some DGPs over others, but that (ii) the rates at which subjects have such affinity varies across DGPs. Call this the "affinity interpretation" of the

[22] In Supplemental Appendix A.3, we use the Bayesian information criterion to reject the hypothesis that a regression model that includes data set dummies improves the fit over a model that does not.

[23] We caution, however, that there are considerably fewer observations per data set than per DGP, which reduces power for this analysis.

variation we observe in extraction rates across DGPs. Under this interpretation, subjects specialize—some subjects may, for instance, be good at detecting switch processes while others may be good at detecting instruct processes.

To distinguish between these interpretations (and avoid committing the "ecological fallacy" in interpreting the data),[24] we look for evidence in the Unique treatment of an ordering that is consistent with the complexity but not the affinity interpretation of the extraction data. Under the complexity interpretation, having extracted a relatively "difficult" DGP (a DGP with a low overall extraction rate) will predict extraction of on-average easier DGPs, while having extracted an "easy" DGP (a DGP with a high extraction rate) should be much less predictive of overall extraction rates. In Figure 2, we plot on the x-axis the average difficulty (across subjects) of extracting each of our eight DGPs (one minus the extraction rate pictured in Figure 1), and on the y-axis, for each DGP, the average rate of extraction (over all DGPs) for the subset of Unique subjects who extracted that DGP.

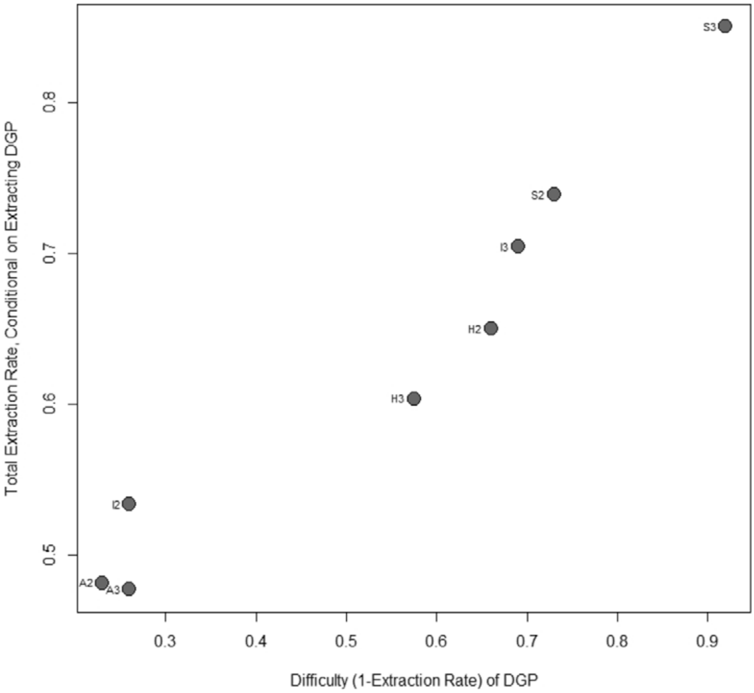


FIGURE 2. Total rate of extraction conditional on extracting a model of each individual DGP. *Notes*: Data is from 792 choices in the Unique treatment. The x-axis plots the difficulty (one minus the extraction rate) of extracting each DGP, calculated across subjects. The y-axis plots, for each DGP, the average extraction rate (over all DGPs) for the subset of subjects who successfully extracted that DGP.

[24] This is the fallacy of supposing that patterns that arise in the aggregate are driven by the same patterns at the individual level.

Under a complexity interpretation, we expect to observe a strong upward sloping relationship, while under the affinity interpretation we need not see any relationship at all. The data clearly shows a strong upwards relationship,[25] indicating a consistent ordering and, therefore, supporting a complexity interpretation of the extraction data.

Result 3. *There is a consistency in extraction behavior across subjects that suggests a complexity ordering across DGPs.*

Figure 3 uses response times across DGPs as additional evidence in favor of our interpretation of the difficulty of extraction as an outgrowth of DGP "complexity." Response times are widely used as measures of the complexity of decision tasks in economics and psychology (e.g., Rubinstein (2007), Gill and Prowse (2023), and Frydman and Jin (2021)). Figure 3 plots the extraction rate for each DGP in each treatment (y-axis) against the median time subjects spend making their full set of decisions in part 2 of the experiment

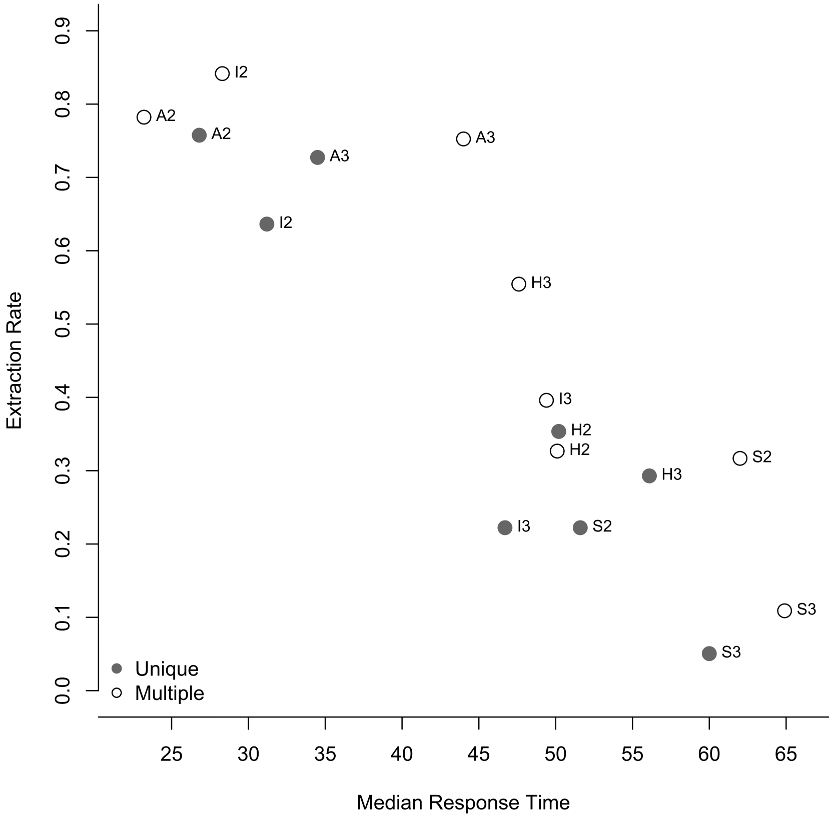


Figure 3. Extraction rates vs. median response time across DGPs. *Notes*: Data is from 1600 choices in the Unique and Multiple treatments. The x-axis plots the median time spent by subjects in part 2 of the experiment. The y-axis plots, for each DGP, the average extraction rate (over all DGPs) for the subset of subjects who successfully extracted that DGP.

[25] The Pearson's correlation coefficient is 0.97, which is highly statistically significant ($p < 0.001$).

(whether they successfully extracted or not). The plot reveals a strong linear relationship and an extremely tight correlation ($\rho = -0.88$, $p < 0.001$) between extraction rates and response time across DGPs. Subjects spend over twice as long on DGPs with the smallest extraction rate as they do on DGPs with the largest extraction rate.[26]

Result 4. *Subjects spend more time attempting to form models of DGPs that they have more difficulty successfully modeling.*

4.3 *Selection*

Our design allows us to study not only extraction, but also "selection": how people choose between data-consistent models when more than one model is capable of rationalizing a data set. In the Unique treatment, we assigned subjects data sets that can be rationalized by only one model (among those describable as fully connected automata with three or fewer states), while in the Multiple treatment we deliberately assigned subjects data sets (generated by the same DGPs as in Unique) that can be rationalized by *multiple* models in this class. Comparing behavior across the two treatments allows us to answer some first questions about the nature of selection.

Figure 1 gives us some first evidence on the impact of selection on subjects' inference. For each DGP, we plot (using large hollow dots) the rate at which subjects in the Multiple treatment extracted *any* (of the many) models that are data-consistent (that can rationalize the data set given in Stage 1 in the Multiple treatment). Comparing the hollow dots (extraction rates from Multiple) to the dark gray dots (from Unique), we find that subjects sometimes (but not always) extract at a higher rate when more than one model is available. Having multiple models available to rationalize a data set overall increases subjects' ability to extract *some* consistent model from that data set (chi-squared test, $p < 0.001$). Conducting chi-squared tests at the level of the true DGP, we find that subjects extract significantly more often in Multiple than Unique in three tasks: H3, I2, and I3 (we cannot reject the null at the 5% level for the other DGPs).

Result 5. *Subjects extract data-consistent models at a higher rate when multiple models are consistent with the data.*

Figure 1 also plots small hollow dots showing the rate at which subjects in Multiple extracted the, ex post, *true* DGP.[27] It is clear that the higher rate of extraction in Multiple is *not* driven by higher rates of extraction of the true DGP. Focusing only on extraction of the (ex post) true model, the rates of extraction are identical in the Unique and Multiple cases (41% in both cases). The increase is therefore due to the availability of models other than the, ex post, correct one.

[26] The direct relationship we observe between response time and mistakes rate is predicted by recently developed"sequential sampling models" of decision-making (e.g., Fudenberg, Strack, and Strzalecki (2018)).

[27] Note that failing to extract the true DGP in the Multiple treatment is *not* irrational. Extracting any data-consistent model based on the part 1 data set is perfectly rational given the information available to subjects. However, ex post, failing to extract the true DGP will cause the subject to make mistakes in part 2.

These patterns suggest that the increased rate of model extraction in Multiple occurs due to something akin to search—when more models are available to rationalize the data set, extraction is more likely to be successful, mirroring the comparative static of any costly search model. This raises the question of whether this process resembles random, undirected search or whether, rather, it is more akin to directed search. In other words, is the increased extraction rate in Multiple a statistical effect of there being more options to stumble upon, or is it driven by the appearance of specific additional models that are easier or more appealing to extract than the true one?

We can test for evidence of undirected search because we deliberately varied the number of rationalizing models available across tasks. For instance, for our S2 task, we gave subjects a data set rationalizable by only 2 models, while for I2 subjects were given one rationalizable by 20. Most tasks featured data sets rationalizable by 3–6 models. Under the hypothesis of random search, there should be a strong relationship between the number of available models and the degree to which extraction improved relative to the same DGP under Unique. We find no such relationship. The Spearman correlation coefficient of 0.34 is not significantly different from zero ($p = 0.41$). The Multiple task with the most data-consistent models available (I2) featured only the third largest improvement over Unique and the task with the largest improvement (H3) had only four data-consistent models. We conclude the following.

Result 6. *The number of models available to rationalize the data set does not predict improvements in extraction rates in the Multiple treatment. This suggests that characteristics of the specific set of data-consistent models available drives these improvements and, therefore, that subjects engage in some form of directed search.*

In Section 5, we structurally estimate the models subjects extract in the Multiple treatment and examine in greater depth what characteristics of models govern selection and this directed search process.

4.4 *Articulation*

Our measure of successful model formation so far has been based exclusively in what we call “extraction”: the rate at which subjects act “as if” they use a model in part 2 that rationalizes the data observed in part 1. Extraction requires only *implicit* learning, relying on subjects’ ability to imitate the pattern of the part 1 data generating process. Do subjects also have an *explicit* understanding of what they have learned in the sense that they can articulate the model explicitly? This need not be true—it is possible that subjects can form a mental model of the true DGP without having the ability to describe what it is they have learned. Indeed, philosophers of science such as Karl Polanyi (Polanyi (2009)) have emphasized that a great part of our knowledge is “tacit knowledge”—knowledge that the learner is unable to describe or articulate in words or symbols and perhaps even has acquired unconsciously. Likewise, psychologists (Reber (1967)) have long emphasized and studied a distinction between “implicit” learning (producing knowledge agents cannot describe and may be unaware of) and “explicit” learning (producing knowledge agents are fully aware of).

Part 3 of each of our tasks allows us to evaluate the rates at which subjects are fully conscious of what they have learned in part 1 and applied in part 2. Recall that, at the end of each task, we ask subjects to describe in words the rule the computer used to produce the outputs in part 1, and we incentivized them to give an accurate and complete description. After data collection, we created a that did not contain the true model or the subject's choices and, line by line, translated the rule the subject wrote into automata descriptions. We also recorded when subjects admitted they were unable to describe or articulate the rule, when they described processes outside of the true class of DGPs, or when they described overly elaborate DGPs that could not rationalize the data. Finally, we matched this data back to the original data set and coded a subject as having successfully articulated the model if the automaton formed from her words fully matched a data-consistent automaton.

Subjects' articulation efforts are often successful. For instance, one subject, in trying to describe her model of I3 wrote, "A double bb results in a y. An a or single b is an x" and another trying to articulate S3 wrote "a output stays consistent with the previous output letter. b output goes in the pattern of yxxyxxyxxyxxyxx..." Often, however, subjects have difficulty articulating models they have successfully extracted. Sometimes subjects admit (often with frustration) that they cannot describe the model they have just successfully extracted, like this subject attempting to articulate H3: "I'm sorry, you're on your own with this one. There's gonna be more x's than y's. I don't understand the pattern very well." Sometimes subjects who successfully extracted express doubt that there actually is a coherent DGP ("I'm not sure there is a rule," or "no idea, seems random"). And sometimes subjects articulate incomplete models, overly elaborate models, or models containing errors. For example, after successfully extracting, one subject attempted to articulate H3 by writing "the letter b input will always produce an x output. The letter a input will produce a y output randomly, but only if there are two or more a inputs in a row. If there is only one a input in a row (e.g., abab), then the output for a is x. The y output is random, but can only be triggered if there are two or more a inputs in a row (e.g., baab)."

Our main question is how rates of articulation (part 3) compare to rates of extraction (part 2)—a comparison that tells us how much of what subjects infer in forming models is implicit. In Figure 4, we plot the mean extraction rate for each model on the x-axis and the corresponding articulation rate on the y-axis, using data from both treatments. The 45-degree line is shown as a dashed line.[28] Virtually, all DGPs appear below the 45-degree line, indicating that subjects are generically worse at articulating models than at extracting them. Overall, subjects articulate at 65% of the rate they extract—a rate that falls to just over 50% for 3-state DGPs. There is significant variation in this proportion. For some DGPs, subjects articulate at rates as low as 28% of the rate at which they extract. This gap between extraction and articulation is highly significant ($p < 0.001$, paired Wilcoxon test on rates of extraction and articulation).

[28] In order to filter out subjects who were undermotivated to articulate in part 3, we remove from the data set subjects who failed to *ever* articulate a model correctly (the results are similar if we include the entire data set instead).

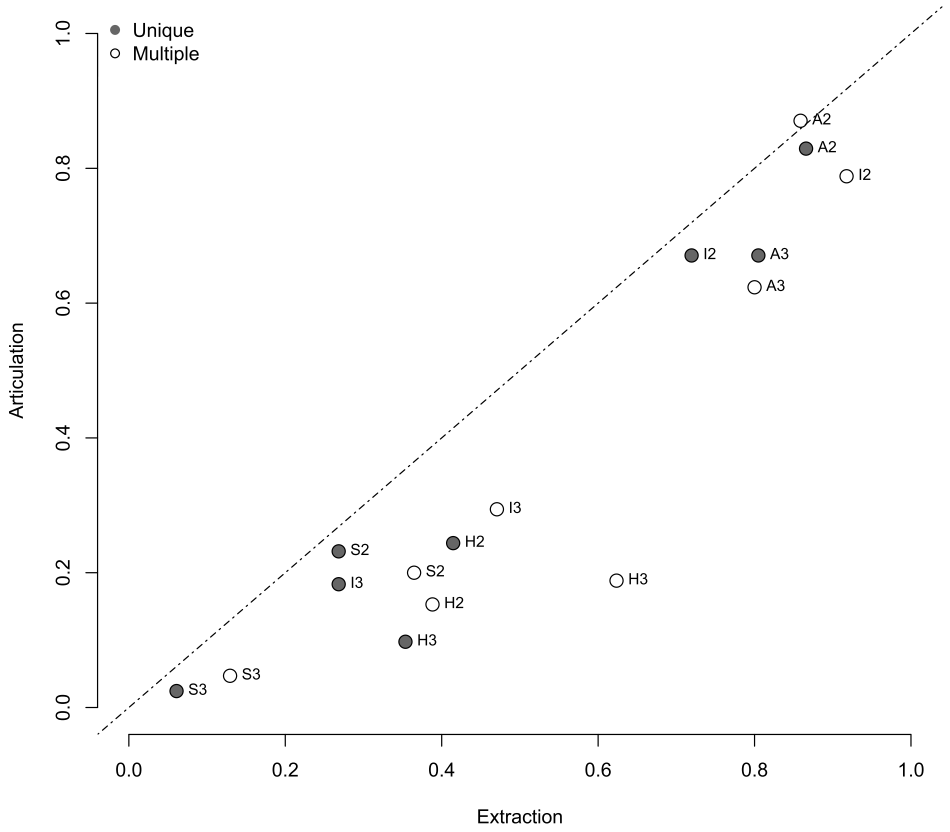


FIGURE 4. Rates of articulation and extraction for each DGP and treatment. *Notes*: Data is from 3200 choices made in the Unique and Multiple treatments. The x-axis plots the extraction rate for each DGP/treatment and the y-axis plots the corresponding rate of articulation of a data–consistent model of each DGP.

Thus, overall, and for some DGPs in particular, much of what subjects learn from inference in part 1 is coded as tacit or implicit knowledge—knowledge that subjects cannot articulate and may even not be explicitly aware of.

RESULT 7. *Subjects are often unable to explicitly articulate the models they extract. This suggests that a great deal of model formation occurs via implicit rather than explicit learning.*

To better understand failures to articulate and to check whether or not subjects are extracting more complex models than we consider when declaring that a subject "extracted" a model, in Figure 5 we categorize and plot the types of responses subjects give in part 3 of the experiment. Overall, we find that most subjects make serious efforts to articulate: nearly 85% of subjects give at least partly coherent advice on at least one task. Thus, the vast majority subjects are willing and capable of articulating models in principle. The left-most category in the figure ("Any Articulation") shows that subjects at least

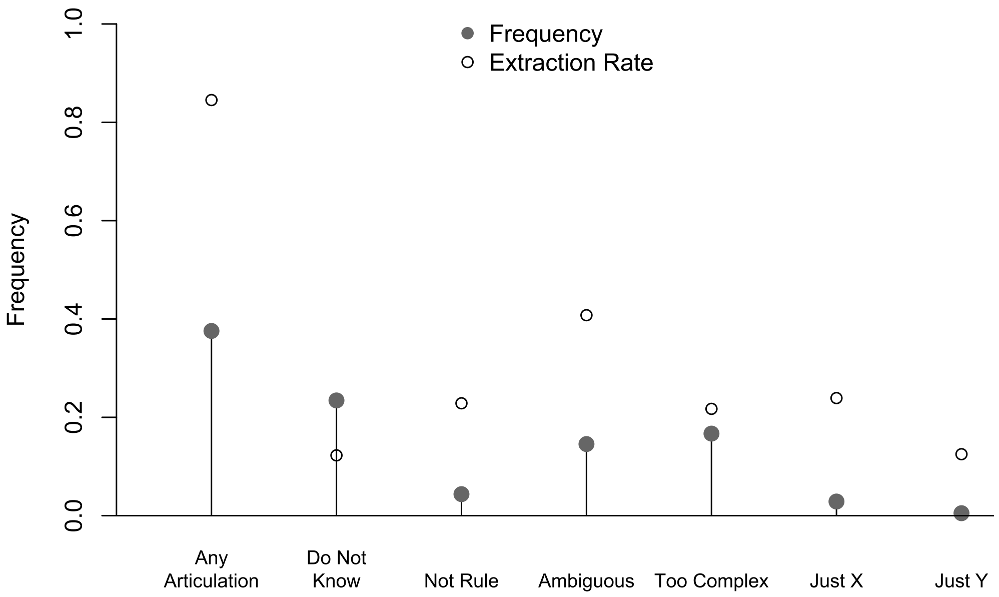


Figure 5. Breakdown of types of responses given in part 3. *Notes*: Data is from 1336 choices in the Unique and Multiple treatments. Dark dots show the frequency of each type of response. Hollow dots show the extraction rate for subjects who responded in each way.

partially articulate the true model just under 40% of the time. The next most common response is the subject saying explicitly that she does not know or cannot articulate what the true model is (23%). The remaining subjects provide descriptions of the model that are not in the form of a rule ("Not a Rule," 4%), too ambiguously described to form an automaton description ("Ambiguous," 15%), present an overly complicated description of the model that fails to rationalize the part 1 data set ("Too Complex," 17%) or give a heuristic prescription to simply choose the same action throughout the task ("Just X/Y").

This pattern of findings is particularly valuable because it directly supports and confirms our operationalization of extraction. Recall, in determining whether subjects extract a model, we restrict attention to models/DGPs with fewer than four states. Our analysis of the articulation data suggests that few subjects (17%) actually form models more complex than this, and almost all of these overly elaborate models are incomplete and fail to rationalize the part 1 data set. Thus, our restriction to 2- and 3-state models in operationalizing extraction is not binding. The fact that only 17% of descriptions are too complex to describe in terms of an automaton with at most three states, provides reassurance that we can restrict to this class of models when deciding whether or not a subject successfully extracted any model. Further support comes from the fact that, of these 17%, almost all only partially describe an automaton with more than four states so that very few indicate successful formation of a more complicated model.

Hollow dots above each of these categories reveal the rates at which subjects who provided each description type managed to extract the true model in their choices in part 2. Subjects who managed to articulate were by far the most likely to extract. But there is also significant extraction for subjects who failed to articulate the true model. Once again, this suggests that many subjects who are unable or unwilling to explicitly articulate the true model, nonetheless were able to implicitly understand the model well enough to use it to guide their actions.

5. What makes a model complex?

Our results illustrate significant variation in the rates of extraction, selection, and articulation of models across DGPs and (to a much lesser extent). In this section, we search for a principled explanation for this variation by examining a number of distinct notions of complexity, drawn from computer science, information theory, algorithmic information theory, and economics. Our aim in this search is (i) to provide some insight into why model formation problems are complex for humans and (ii) to examine whether there is a unity in the notions of complexity that drive the somewhat different cognitive acts of extraction, articulation, and selection.

5.1 *Complexity notions*

We collected a total of 14 notions/measures of complexity that plausibly predict the complexity of forming models and group them into three broad categories, described in the following three subsections. Additional details for partition complexity and sparsity complexity, as well as the algorithms used to calculate them, are provided in Supplemental Appendix B. Notably, although some of these complexity metrics (s-complexity, t-complexity) are related to the automaton descriptions of DGPs provided in Table 1, most are built on orthogonal descriptions that are unrelated to automata.

5.1.1 *Implementation notions* First, the complexity of forming a model may be linked to characteristics of the DGP itself, and the costs of implementing it. In particular, there are several notions advanced in economics and computer science of what Oprea (2020) calls "implementation complexity"—the resource burden of implementing the DGP's automaton for a human or a computational device. We consider three salient measures in this class, which we collectively call **Implementation Measures**.

States (s-complexity): The number of states in the automaton describing the DGP. This is the most popular measure of implementation complexity in the automata literature in economics (e.g., Rubinstein (1986)) and was found to be an important driver of complexity costs in Oprea (2020).

Transitions (t-complexity): The number of transitions in the automaton describing the DGP, hypothesized to generate complexity costs by Banks and Sundaram (1990) and verified as a significant source of complexity costs in Oprea (2020). This is the number of arcs in the automaton diagrams shown in Figure 1.[29]

Machine Complexity: The size of the DGP as measured by the minimum number of two-input gates (NOR, NAND, etc.) that would be required to implement it physically in

[29] Note that our transition complexity measure is different from the one used in Oprea (2020), reflecting our very distinct question. Oprea (2020) focused on a measure of transitions in excess of states to separately identify the contribution of states and transitions to implementation costs.

hardware. To the degree human brains work similar to physical silicon chips, machine complexity describes the cognitive effort required to implement a correct model of the DGP for prediction.

5.1.2 *String/sequence notions* Second, at the other extreme, are measures focused entirely on characteristics of the data set (rather than the DGP). In particular, these **string/sequence measures** focus on characteristics of the input or output strings that constitute the data set subjects observe in part 1.

Entropy: The Shannon entropy of the input string or output string in the data set. This is a measure of the amount of information or surprise in a string and is calculated as $-\sum_{u \in U} p(u) \log(p(u))$ (for an input string), where $p(u)$ is the probability of u (approximated by the relative frequency in the string). We separately consider the entropy of inputs (**entropy in)** and the entropy of outputs (**entropy out**) as two different measures.

Kolmogorov complexity: A measure of the computational resources needed to create a string (e.g., the input or output string): the length of the shortest computer program that can create the string. Intuitively, it characterizes a string or sequence as being easier to draw inferences from if it features stronger regularity. It is the core complexity notion in algorithmic information theory, but an important set of theorems (see Li and Vitanyi (2008)) show that it is uncomputable. However, computable approximations to Kolmogorov complexity exist and we include these as complexity measures. One approximation is *Lempel–Ziv complexity* (Lempel and Ziv (1976)), which we calculate separately for inputs and outputs (**LZ in** and **LZ out**). Another is *approximate entropy* (Pincus, Gladstone, and Ehrenkranz (1991)) (**approximate in** and **approximate out**).[30,31]

5.1.3 *Relational notions* Finally, we consider measures that focus on both the data set and the DGP. These **relational metrics** are descriptions of the relationship between input and output strings generated, on average, by the DGP.

Mutual Information: A statistical measure of the symmetric informativeness of one random variable about another that is widely used as a measure in rational inattention models Sims (2003)). We approximate the mutual information between the entire set of inputs, $u^T \equiv \{u_t\}_{t=1}^{T} \in U^T$ and the current output, v_T, by simulating the DGP for 5000 periods to obtain an approximation of the joint distribution, $P(v_t = v, u^T = \tilde{u})$ with $\tilde{u} \in U^T$. Mutual information is then calculated as $I(u^T, v_T) = \sum_{v,u} P(v_t = v, u^T = u) \log \frac{P(v_t=v, u^T=u)}{P(v_t=v)P(u^T=u)}$, where T is the size of the data set a decision maker has to make inferences from.

Sensitivity: A deterministic measure of the sensitivity of outputs to the history of inputs in DGPs, proposed by Lipman and Srivastava (1990). Specifically, the measure counts the number of alterations one can make to the history of inputs that would

[30] Approximate entropy was developed for measuring regularity or irregularity in heartbeats. Lempel–Ziv complexity was developed by computer scientists as a practical alternative to Kolmogorov complexity that forms the basis for Lempel–Ziv lossless compression.

[31] More promising might be conditional Kolmogorov complexity, which is measured as the length of the shortest computer program that can create the output string *from the given input string*. Unfortunately, this too is uncomputable and we are not aware of any methods for approximating it.

change the output at the end of that history, summed over all histories.[32] Our main measure calculates this at the DGP level, for example, over all possible histories. We will call this notion simply **sensitivity**. We also calculate a version fitted to the specific part 1 subjects observed, which we call **string sensitivity**.

Partition complexity: A measure of the information processing of the data set required to correctly implement the DGPs output sequence, adapted from a framework for modeling bounded rationality suggested by Lipman (1995). Our adaptation counts the number of unique elements in the coarsest partition of the set of subhistories in the data set that a decision maker must use to correctly forecast the outputs required by the DGP. The idea is that models that require an agent to distinguish between more distinct patterns in the data set (more unique partition elements) are more difficult to identify and model. For instance, a model of DGP I2 only requires the agent to partition the data set into two kinds of elements to forecast outputs: (i) output x occurs if the input is a and (ii) output y occurs if the input is b. But DGP S2 (which has the same number of states and transitions) requires the subject to distinguish between four patterns: (i) output x occurs if the input is a and previous output is x; (ii) output y occurs if the input is a and the previous output is y; (iii) output x occurs if the input is b and the previous output is y; (iv) output y occurs if the input is b and the previous output is x. Thus, S2 is more "partition complex" than I2.

Sparsity complexity: A measure of the sparsity of a DGP—the amount of the data set that can be safely ignored when applying a correct model of the DGP to predict outputs. The measure is an adaptation of ideas from Gabaix (2014), which postulates that sparser models are less difficult or costly to apply. To operationalize, we say that a model/DGP is "less complex in the sense of sparsity" if fewer elements (inputs or outputs) in the data set need to be attended to in order to predict outputs. For instance, S2 is more "sparsity complex" than I2 because an agent must attend to both the input and previous output in S2, but only the input in I2. This measure is closely related to partition complexity and, like partition complexity, is a measure of the amount of information in a data set that must be processed to apply the model.[33]

5.2 *The complexity of extraction and articulation*

In Figure 6, we plot for each complexity notion, the absolute value of the estimated correlation between the complexity metric and (i) extraction (the y-axis) and (ii) articu-

[32] We weight all histories equally and only consider alterations in which one input changes. Lipman and Srivastava (1990) consider a more general model.

[33] One additional notion we hoped (but were unable) to use is "perceptrons," which have been successfully used in modeling bounded rationality in economics (e.g., Rubinstein (1993)). Perceptrons are binary classifiers that produce an output classification by calculating the weighted sum of a set of inputs and comparing it to a threshold. The number of inputs (a perceptron's order) can be taken to be a measure of its complexity (Rubinstein (1993)). Thus, a natural measure of the complexity of a model is the minimal order of the perceptron that can implement it. However, in attempting to adapt it we ran into a well-known problem of perceptrons—they can only implement linearly separable functions. If the function corresponding to a model is not linearly separable, no perceptron exists. For example, the output of S2 corresponds to the XOR of the previous input and output, which is the quintessential example of a function that is not implementable by a perceptron. This nonimplementability prevents us from being able to use perceptrons to construct a complexity measure.

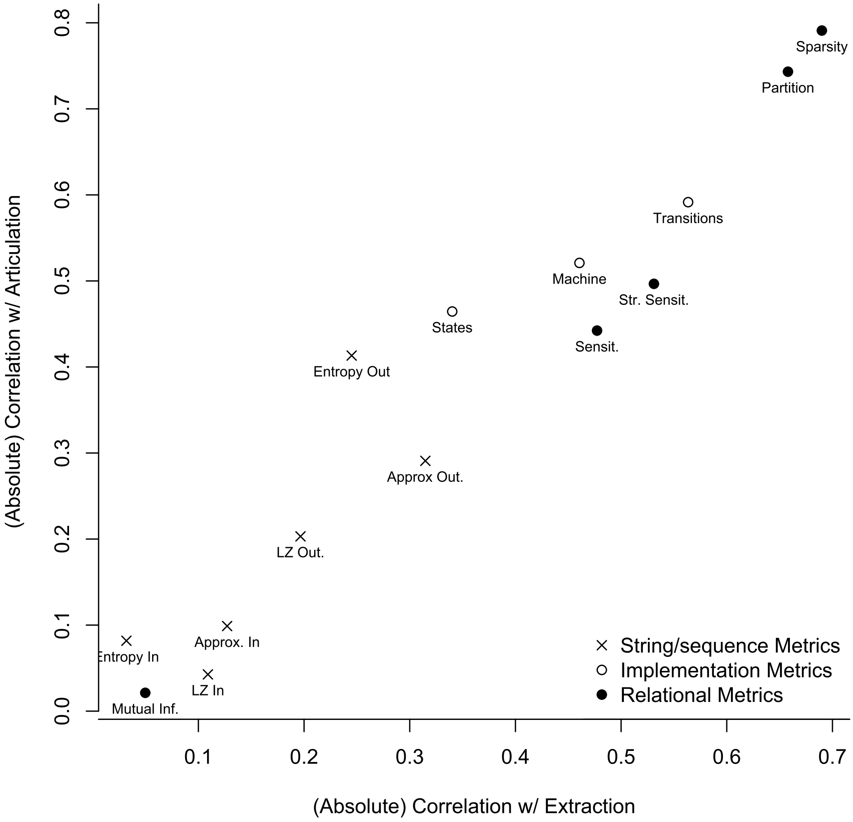


Figure 6. Absolute value of the correlation between complexity metrics and articulation/extraction. *Notes*: Data is from 1584 decisions in the Unique treatment. A separate dot is presented for each complexity notion/metric. The absolute value of Goodman/Kruskal gamma correlation is plotted for extraction on the x-axis and for articulation on the y-axis.

lation (the x-axis), using data from the Unique treatment.[34] To complement, in Supplemental Appendix A, we show scatterplots of mean extraction and articulation rates against each of these 14 measures.[35]

We make three observations.

First, there is an extremely strong relationship between the values of complexity metrics in explaining each of extraction and articulation. High correlation of a metric with one measure of model formation tends to be accompanied by high correlation with the other.

[34] For this calculation, we use Goodman and Kruskal's gamma, which like Spearman correlation and Kendall's tau, measures the rank correlation between two variables. The main advantage of Goodman and Kruskal's gamma is that it allows for more general relationships in the data than Spearman and does not penalize ties like Kendall's tau (and thus does not penalize the metrics that only have a few values such as states).

[35] Most of the correlations plotted in Figure 6 are statistically significantly different from zero at the 5% level. Only mutual information, LZ in, and approximate in are not.

Second, string/sequence measures like entropy and approximations to Kolmogorov complexity (LZ In/Out and Approx In/Out) tend to be particularly bad predictors of model formation. This mirrors and reinforces the observation from our analysis of extraction that DGPs are a much stronger driver of variation than are data sets. Indeed, implementation metrics (states, computational complexity, and transitions) that are rooted in the DGP are all better than the string/sequence measures.[36]

Third, there is significant variability in explanatory power across Relational measures. The overall worst measure (mutual information) is a relational measure, but so are the very best (partition complexity and sparsity complexity).

Most importantly, we find that two closely related metrics that describe the *information processing* required to apply a model for prediction—partition complexity and sparsity complexity—do the best job of explaining what makes model formation complex, with impressive correlations of 2/3 or more.[37]

Result 8. *Measures of the amount of information processing of the data set required strongly predict subjects' abilities to extract and articulate the model.*

5.3 *Complexity and selection*

Next, we use these same complexity notions to return to the question of what drives selection of models when multiple models are available. Recall from Section 4.3 that (i) subjects tend to extract at a greater rate when multiple models can rationalize the data set, but (ii) the sheer number of available models is a poor predictor of this improvement in extraction across tasks. Our conclusion from this analysis is that selection is likely driven by something akin to directed search, with subjects preferentially selecting some models over others based on some characteristics of the underlying DGP.

In this section, we consider the possibility that something like a behavioral analogue of Occam's Razor applies—that subjects preferentially select models that are simpler under some notion of simplicity. To do this, we estimate a structural model described in Supplemental Appendix A.2 on part 2 choice data to determine which of the available data-consistent models subjects choose in the Multiple treatment. Then we test the hypothesis that subjects are drawn to simplicity (and examine what sort of simplicity they

[36] To give string-based metrics their best shot, we also calculated correlation coefficients within-DGP, effectively using them only to explain the small and somewhat inconsistent variation we observe across data sets. Thus, we calculated correlation coefficients between our string-based complexity metrics (entropy in, entropy out, LZ in, LZ out, approx. in, approx. out, and string sensitivity) and extraction/articulation *within* DGP for each DGP. Although for some DGPs correlations are quite strong (rising to as high as 0.77 for some DGP/metric combinations), we find little systematic evidence that string-based metrics do much better within than between DGPs. Averaging within-correlations across DGPs, we find they are never higher than 0.46 for any metric. We conclude that the variation we observe across data sets is either due to sampling error or to perceptual difficulties presented by some data sets that are ill-captured by our complexity metrics.

[37] While we cannot statistically distinguish the predictive power of partition complexity and sparsity complexity from one another, we can distinguish each of these measures from all of the others. For both of these measures (i) 95% confidence bands on estimates do not overlap estimates from any of the other measures and (ii) confidence bands for sparsity/partition complexity either do not overlap or only slightly overlap confidence bands around the estimates for other measures.

are drawn to) by examining how frequently subjects select models of the lowest complexity under each of our complexity notions.

We estimate subjects' models from part 2 data using a standard random utility framework, specifically a finite mixture model adapted from the Strategy Frequency Estimation Method (SFEM) developed by Dal Bó and Fréchette (2011) to estimate strategies used in repeated games. This procedure give us individual level estimates of the weight each subject places on each of the models (of fewer than four states) available to rationalize the data sets observed in part 1. We use these subject/task level estimates to summarize which models subjects are drawn to in the following way. For each task in the Multiple treatment, we identify the *simplest* (least complex) available data-consistent model *under each complexity notion*. Because this exercise concerns models and not data sets, we use only complexity notions that relate purely to characteristics of the DGP: sensitivity, mutual information, states, transitions, machine complexity, partition complexity, and sparsity complexity.

Next, we calculate the fraction of subjects who clearly selected this simplest model under each metric. For this, we consider only subject/task combinations in which there is clear evidence that the subject extracted a particular model: subject-tasks in which the subject put a weight greater than 0.95 on one of the models according to the structural estimates. These cases make up 58% of the data. Our question is whether and under what complexity notions these subjects put greater than random weight on the simplest available model.

Figure 7 plots the rate at which subjects selected the simplest model as solid lines with dots for each of our complexity notions. A dotted horizontal line shows the rate at which subjects would select the simplest model if they were choosing randomly.

The results show that under *any* complexity notion, subjects select the simplest model more often than random. Moreover, as with extraction and articulation, selection is most strongly explained by the information processing notions (partition complexity and sparsity complexity). Subjects choose the simplest model under these metrics more than 2/3 of the time (almost three times as often than would be expected from random selection), suggesting that subjects are not only particularly good at extracting models with low information processing requirements; they are also strongly attracted to (or good at discovering) these same types of models. These findings therefore suggest a notable unity in the determinants of model formation.

Result 9. *Subjects have a "bias" toward simplicity in selecting models when multiple models are consistent with data. As with extraction and articulation, the strongest predictors of selection are information processing notions of complexity* (*partition complexity and sparsity complexity*).

Finally, we emphasize that the preceding analysis is deliberately conservative and should be interpreted as a lower bound estimate of the weight subjects attach to simple models in selection. In calculating this weight, we are sometimes confronted with the fact that there is no unique simplest model—often multiple available models are equally simple. In these cases, we averaged the weight across that placed on all of the simplest

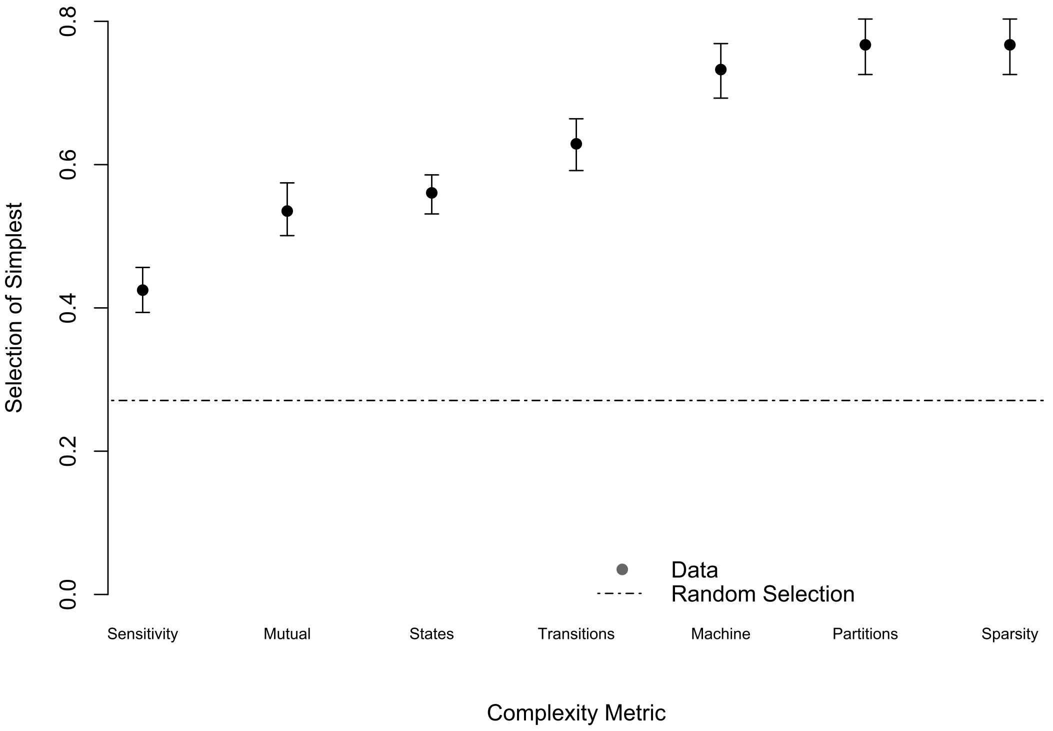


Figure 7. Proportion of subjects who select the simplest available model under various notions of simplicity. *Notes*: Data is from 808 decisions in the Multiple treatment. Each data point corresponds to a different notion of complexity. The y-axis shows the proportion of subjects who selected the most commonly selected of the simplest available model under each notion. The error bars indicate bootstrapped 95% confidence intervals.

models. A less conservative approach would have been to add up the weight across all models in the simplest set—an approach that mechanically overweights complexity notions that make fewer distinctions across models (that have fewer distinct values).

6. Discussion

In the last half century, economists have learned a great deal about the difficulties people face in making inferences. Most of the hundreds of empirical investigations on inference in economics center on the difficulties people face in extracting simple probabilities from stochastic reservoirs of information (see, e.g., Benjamin (2019)). But in economic life, many of our most important inference tasks involve extracting, not simple probabilities, but rather *algorithms* from data. The data generating processes we have to understand in order to forecast and choose, after all, often have rule-like structures ranging from very simple cyclical regularities to elaborate causal relationships. Forming models of such processes to inform beliefs and guide behavior is difficult, not due to stochasticity (as in simple probabilistic inference) but rather to the sheer computational difficulty of extracting patterns from data and representing these patterns in form suitable for prediction.

Our experiment takes some first steps toward understanding how humans form models of algorithms and what makes this type of inference complex (costly or difficult). Because these are first steps, we radically simplify the problem on multiple dimensions. In order to isolate the effects of computational difficulty, we study perfectly deterministic data generating processes. Likewise, we study processes in which the direction of causality is clear to decision makers, ex ante. Finally, we study the simplest possible examples from this already simple class of problems: DGPs describable by 2- and 3-state finite automata with data sets consisting of binary inputs and outputs. All of these simplifications can be relaxed by applying our methods in future work.

Despite these simplifications, we find evidence that forming accurate models of algorithms is extremely difficult for decision makers in our experiment. Although our data generating processes (DGPs) are rudimentary, deterministic, and causally clear, subjects are able to form accurate models of them less than half of the time. This difficulty varies dramatically across DGPs, with subjects accurately forming models of some DGPs upward of 80% of the time and of others less than 10% of the time.[38]

Importantly, we find significant structure to this complexity and are able to make progress in identifying its source. We find evidence of a consistent (across subjects) complexity ordering across DGPs: subjects who successfully form on-average difficult models typically are also able to form on-average easier models. Searching over a wide array of formal complexity notions, we find that this complexity ordering is best organized by a pair of closely related metrics of the information processing required to form a model of the DGP. One of these notions ("partition complexity"), adapted from Lipman (1995), counts the number of distinct patterns a modeler has to identify in a data set in order to make future predictions. Closely related is "sparsity complexity," adapted from Gabaix (2014), which counts the number of distinct input/output elements in the data set a modeler has to attend to in order to predict. Both of these measures of information processing are highly correlated with rates of extraction across modeling tasks.

These findings put some important context on the often remarked-upon fact that decision makers seem eager to borrow and reuse models from other people and contexts rather than form new models of their own. For instance, human beings seem eager to borrow models from their social environment, adopting narratives and explanatory schema provided by family members, peers, mentors, and political and religious affiliates. This social and cultural contagion of models is understandable in light of the difficulty people seem to have in forming even simple models in our data on their own. Likewise, decision makers often seem to use thin analogies between circumstances as bases for reusing mental models from very different contexts, even when these secondhand models are not perfectly appropriate fits. It is plausible that one of the roots of

[38] Similarly, we should expect the rate at which decision makers fail to form accurate mental models to vary with context, incentives, and opportunity costs of time. For this reason, we view our findings regarding the predictive power of complexity metrics across DGPs as our primary contribution, rather than our documentation of the levels of failure observed in our particular experimental sample.

the widespread use of heuristics in decision-making observed in behavioral economics and psychology is the sheer cost and difficulty people face when attempting to form new models of novel problems. The fact that this complexity has significant structure suggests we may be able to model such phenomena and predict the contexts in which social contagion and heuristic reuse of models will tend to arise.

In addition to providing evidence on how well subjects "extract" models to guide behavior, our experiment also allows us to examine how well they can consciously "articulate" what they have learned. We find that subjects are often able to extract models that they are not able to afterwards accurately describe. This is evidence that at least some of the learning involved in modeling complex phenomena is implicit and perhaps even unconscious. Nonetheless, we find that the same formal notions of complexity—ones that measure the information processing required to recognize an algorithm—best predict articulation just as with extraction.

Finally, our design allows us to understand not only what makes an algorithm difficult to accurately model, but also what types of models people are preferentially drawn to when explaining data. When more than one model is available to rationalize data in our experiment, we find a kind of behavioral analogue to Occam's Razor: subjects tend to preferentially select simple models to explain data. What is more, this model selection is governed by the same complexity notions—partition complexity and sparsity complexity—that govern the difficulty of extracting and articulating models, with subjects preferentially selecting models that require less information processing. The fact that selection is "biased" in this way may be important for understanding bounded rationality and its persistence. A growing theoretical literature on misspecified models in economics shows that mistaken models are often resistant to learning, leading to persistent and even reinforcing mistakes in behavior and beliefs. Our findings point to structure in what types of models people may tend to get stuck in and, therefore, may be useful for building future models of bounded rationality.

Our results reveal a fundamental unity in the complexity of forming models. Extraction, articulation, and selection of models are all best organized by the same metrics of complexity, rooted in the information processing required to form and deploy a model. Future work should test the limits of this unity. Does the same type of complexity organize the formation of models of DGPs complicated by stochasticity (e.g., nondegenerate Markov chains) or confounded by causal uncertainty (e.g., DAGs)? Conducting experiments that extend our methods to these richer settings seem an important next step in this agenda.

More generally, our findings suggest that the difficulty of finding a solution to a problem (e.g., inferring a correct model) is directly linked to the information processing required to later deploy that solution in decision-making (e.g., to *use* a model to make predictions). Is this true more generally of the wide range of judgemental and optimization failures documented in behavioral and experimental economics? Answering this question will require continued development of models that characterize the information processing required of rational behavior, and experiments testing the organizing power of resulting metrics in explaining deviations from optimal choice in a much wider range of settings. Exploring these questions both theoretically and empirically seems a promising path toward learning how to build robust, predictive models of boundedly rational decision makers.

References

Abeler, Johannes and Simon Jäger (2015), "Complex tax incentives." *American Economic Journal: Economic Policy*, 7 (3), 1–28. [180]

Abreu, Dilip and Ariel Rubinstein (1988), "The structure of Nash equilibrium in repeated games with finite automata." *Econometrica*, 56 (6), 1259–1281. [180]

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler (2005), "Fact-free learning." *American Economic Review*, 95 (5), 1355–1368. [179]

Banks, Jeffrey S. and Rangarajan K. Sundaram (1990), "Repeated games, finite automata, and complexity." *Games and Economic Behavior*, 2 (2), 97–117. [199]

Ben-Porath, Elchanan (1990), "The complexity of computing a best response automaton in repeated games with mixed strategies." *Games and Economic Behavior*, 2 (1), 1–12. [180]

Benjamin, Daniel J. (2019), "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations*, 1 (2), 69–186. [176, 205]

Bohren, Aislinn and Daniel Hauser (2021), "Learning with heterogeneous misspecified models: Characterization and robustness." *Econometrica*, 89, 3025–3077. [179]

Byrne, Ruth and Philip N. Johnson-Laird (1989), "Spatial reasoning." *Journal of Memory and Language*, 28, 564–575. [180]

Caplin, Andrew (2016), "Measuring and modeling attention." *Annual Review of Economics*, 8, 379–403. [180]

Chatterjee, Kalyan and Hamid Sabourian (2009), "Game theory and strategic complexity." In *Encyclopedia of Complexity and Systems Science*, 4098–4114. [180]

Dal Bó, Pedro and Guillaume R. Fréchette (2011), "The evolution of cooperation in infinitely repeated games: Experimental evidence." *American Economic Review*, 101 (1), 411–429. [204]

Dean, Mark and Nathaniel Neligh (2023), "Experimental tests of rational inattention." *Journal of Political Economy*, forthcoming. [180]

Eliaz, Kfir and Ran Spiegler (2020), "A model of competing narratives." *American Economic Review*, 110 (12), 3786–3816. [180]

Enke, Benjamin (2020), "What you see is all there is." *Quarterly Journal of Economics*, 135 (3), 1363–1398. [179]

Esponda, Ignacio and Demian Pouzo (2016), "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica*, 84 (3), 1093–1130. [179]

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel (2023), “Mental models and learning: The case of base-rate neglect.” *American Economic Review*, forthcoming. [179]

Frydman, Cary and Lawrence Jin (2021), “Efficient coding and risky choice.” *The Quarterly Journal of Economics*, 137, 161–213. [193]

Fudenberg, Drew, Gleb Romanyuk, and Philipp Strack (2017), “Active learning with a misspecified prior.” *Theoretical Economics*, 12 (3), 1155–1189. [179]

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki (2018), “Speed, accuracy, and the optimal timing of choices.” *American Economic Review*, 108 (12), 3651–3684. [194]

Gabaix, Xavier (2014), “A sparsity-based model of bounded rationality.” *The Quarterly Journal of Economics*, 129 (4), 1661–1710. [175, 178, 180, 201, 206]

Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwarzstein (2021), “Channeled attention and stable errors.” [179]

Gilboa, Itzhak (1988), “The complexity of computing best-response automata in repeated games.” *Journal of Economic Theory*, 45 (2), 342–352. [180]

Gill, David and Victoria L. Prowse (2023), “Strategic complexity and the value of thinking.” *The Economic Journal*, 133 (650), 761–786. [193]

Graeber, Thomas (2023), “Inattentive inference.” *Journal of the European Economic Association*, 21 (2), 560–592. [179]

Handel, Benjamin and Joshua Schwartzstein (2018), “Frictions or mental gaps: What’s behind the information we (don’t) use and when do we care?” *The Journal of Economic Perspectives*, 32 (1), 155–178. [179]

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein (2014), “Learning through noticing: Theory and evidence from a field experiment.” *The Quarterly Journal of Economics*, 129 (3), 1311–1353. [179]

Heidhues, Paul, Botond Koszegi, and Philipp Strack (2018), “Unrealistic expectations and misguided learning.” *Econometrica*, 86 (4), 1159–1214. [179]

Jimenez, Luis, Joaquin Vaquero, and Juan Lupianez (2006), “Qualitative differences between implicit and explicit sequence learning.” *Journal of Experimental Psychology Learning, Memory, and Cognition*, 32 (3), 475–490. [181]

Jin, Giner, Michael Luca, and Daniel Martin (2021), “Complex disclosure.” *Management Science*, 68 (5), 3236–3261. [180]

Kalai, Ehud and William Stanford (1988), “Finite rationality and interpersonal complexity in repeated games.” *Econometrica*, 56 (2), 397–410. [180]

Kendall, Chad, and Ryan Oprea (2024), “Supplement to ‘On the complexity of forming mental models’.” *Quantitative Economics Supplemental Material*, 15, https://doi.org/10.3982/QE2264. [186]

Lempel, Abraham and Jacob Ziv (1976), “On the complexity of finite sequences.” *IEEE Transactions on Information Theory*, 22 (1), 75–81. [200]

Li, Ming and Paul M. B. Vitanyi (2008), *An Introduction to Kolmogorov Complexity and Its Applications*, third edition. Springer Publishing Company, Incorporated. [200]

Lipman, Barton L. (1995), “Information processing and bounded rationality: A survey.” *Canadian Journal of Economics*, 28 (1), 42–67. [175, 178, 180, 201, 206]

Lipman, Barton L. and Sanjay Srivastava (1990), “Informational requirements and strategic complexity in repeated games.” *Games and Economic Behavior*, 2 (3), 273–290. [200, 201]

Mathews, Robert and Ray Buss (1989), “Role of implicit and explicit processes in learning from examples: A synergistic effect.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15 (6), 1083–1100. [181]

Murawski, Carsten and Peter Bossaerts (2016), “How humans solve complex problems: The case of the knapsack problem.” *Scientific reports*, 6, 34851. [180]

Oprea, Ryan D. (2020), “What makes a rule complex.” *American Economic Review*, 110 (12), 3913–3951. [180, 188, 189, 199]

Pincus, Steven M., Igor M. Gladstone, and Richard A. Ehrenkranz (1991), “A regularity statistic for medical data analysis.” *Journal of Clinical Monitoring and Computing*, 7 (4), 335–345. [200]

Polanyi, Michael (2009), *The Tacit Dimension*. University of Chicago Press. [195]

Reber, Arthur (1967), “Implicit learning of artificial grammars.” *Journal of Verbal Learning and Verbal Behavior*, 6, 855–863. [181, 195]

Rubinstein, Ariel (1986), “Finite automata play the repeated prisoner’s dilemma.” *Journal of Economic Theory*, 39 (1), 83–96. [180, 199]

Rubinstein, Ariel (1993), “On price recognition and computational complexity in a monopolistic model.” *Journal of Political Economy*, 101 (3), 473–484. [201]

Rubinstein, Ariel (2007), *Instinctive and Cognitive Reasoning: A Study of Response Times*, Vol. 117. [193]

Schotter, Andrew and Barry Sopher (2003), “Social learning and coordination conventions in intergenerational games: An experimental study.” *Journal of Political Economy*, 111 (3), 498–529. [186]

Schwartzstein, Joshua and Adi Sunderam (2021), “Using models to persuade.” *American Economic Review*, 111 (1), 276–323. [180]

Shanks, David, Theresa Johnstone, and Leo Staggs (1997), “Abstraction processes in artificial grammar learning.” *The Quarterly Journal of Experimental Psychology*, 50A (1), 216–252. [181]

Sims, Christopher A. (2003), "Implications of rational inattention." *Journal of Monetary Economics*, 50 (3), 665–690. [180, 200]

Spiegler, Ran (2016), "Bayesian networks and boundedly rational expectations." *Quarterly Journal of Economics*, 131 (3), 1243–1290. [180]

Co-editor Garance Genicot handled this manuscript.

Manuscript received 5 October, 2022; final version accepted 28 September, 2023; available online 2 October, 2023.

The replication package for this paper is available at https://doi.org/10.5281/zenodo.8335264. The Journal checked the data and codes included in the package for their ability to reproduce the results in the paper and approved online appendices. Given the highly demanding nature of the algorithms, the reproducibility checks were run on a simplified version of the code, which is also available in the replication package.