PREFERRING SIMPLICITY

Itzhak Gilboa and Larry Samuelson

4-09

June, 2009

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

PREFERRING SIMPLICITY[1]

Itzhak Gilboa[2] and Larry Samuelson[3]

October 11, 2008

Abstract

This paper examines circumstances under which a preference for simplicity enhances the effectiveness of inductive reasoning. We consider a game in which Fate chooses a data generating process and agents are characterized by inference rules that may or may not favor simpler theories over more complex ones. The basic intuition is that agents who do not prefer simple theories to more complex ones are doomed to "overfit" the data and therefore engage in ineffective learning. The analysis places no computational or memory limitations on the agents—a preference for simplicity emerges in the presence of unlimited reasoning powers.

[1]We thank Daron Acemoglu, Ken Binmore, and Arik Roginsky for discussions, comments, and references.

[2]HEC, Paris, Tel-Aviv University, and Cowles Foundation, Yale University. tzachigilboa@gmail.com.

[3]Yale University. Larry.Samuelson@yale.edu.

Contents

1 Introduction

People tend to prefer simpler explanations and simpler theories to more complex ones. This preference for simplicity has been championed on normative grounds (most famously by William of Occam (see Russell [13])) and has long been offered as a descriptive model of human reasoning (e.g., Wittgenstein [22]). In everyday reasoning as well as scientific work, a simpler theory is routinely preferred to a more complex one of equivalent explanatory power.

Why do people prefer simpler theories to more complex ones? Is it only because of bounded rationality? Is the preference for simplicity a psychological bias that leads to erroneous decisions, and that should be corrected?

People undoubtedly sometimes opt for simple explanations at the cost of accuracy, in the process choosing suboptimal theories.[1] But we argue in this paper that there are good reasons to prefer simple theories, apart from any considerations of bounded rationality.

1.1 Simplicity Enhances Learning

Example 1 illustrates our main point: A preference for simplicity leads to more effective learning.

Example 1: Lisa and Sally, adjusting their commuting habits to a new job, check whether the bus runs between their home and the office in consecutive 12-hour periods, observing the pattern

$$0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1,$$

where a 1 denotes that the bus runs. They are now asked their prediction for the seventh and subsequent periods. To make this prediction, they must choose between the many theories consistent with the data they have observed. Sally chooses the simplest theory that matches the data, which, according to her judgment, is "$f(n) = 1$ iff n is even." Her prediction is that the following six observations will be 010101. Lisa, in contrast, exhibits

[1] To paraphrase Einstein (our emphasis), scientific theories should be as simple as possible, *but not simpler*.

no concern for simplicity or any other criterion that might help her choose among the various theories that fit the data, viewing any one as being as good as any other. She selects the theory "$f(n)$ is the nth digit in the binary expansion of $\frac{43}{128}$," and predicts that the next six observations will be 100000. Who is more likely to be disappointed the next evening, when Sally drives to work while Lisa waits at the bus stop? ■

Without Sally's preference for simplicity, Lisa is doomed to helplessly puzzle which, amongst the inevitable multitude of theories that fit the data, should be used for prediction. Indeed, Lisa could as well have taken the expansion of $\frac{42}{128}$ as her theory, predicting the bus would never run again, or $\frac{87}{256}$, predicting it would run the next two periods, or ....

As our next example shows, a preference for simplicity is in fact part of our definition of "intelligence." Someone who can easily conceive of all theories that have not been refuted, but cannot distinguish among them, would predict and behave just like one who can conceive of no such theories at all.

Example 2: Lloyd and Sam take an IQ test in which they are asked to extend the sequence

$$1 \quad 2 \quad 4 \quad 8 \quad \ldots .$$

Sam (like Sally) has a taste for simplicity and selects the continuation 16. Lloyd (like Lisa) has no concern for simplicity or anything else beyond fitting the data, and chooses the value 7.

On the strength of this and similar problems, Sam is likely to be judged very intelligent, whereas Lloyd is likely to be taken for a dolt. However, Lloyd may well protest this assessment. Indeed, Lloyd may be free of any computational limitations and well aware that the sequence is consistent with the function

$$f(n) = 2^{n-1}, \tag{1}$$

yielding the continuation 16. However, Lloyd may have also realized that this sequence is consistent with the function

$$f(n) = -\frac{1}{3}n^4 + \frac{7}{2}n^3 - \frac{73}{6}n^2 + 18n - 8,$$

whose next value is $f(5) = 7$, and which (at least to Lloyd) seems just as good a theory by which to predict the next value as does (1). ■

One might object that what Lloyd lacks is not a notion of simplicity but social intelligence. He evidently does not understand what is expected of him on such an exam. If he had a "theory of mind" of the others around him, he would know that they would think of the first function rather than the second, without having to use any notion of simplicity to sort through theories. However, this would be unfair to Lloyd: drawing such an inference about others, their computational abilities, and their tastes and norms also requires a preference for simplicity. Lloyd might justifiably ask, "How would I know that testers here prefer simpler functions, only because testers in the past have exhibited such a preference?"

An alternative account might similarly suggest that what Sally exhibits is a taste for *plausibility* rather than simplicity. She chose her theory, that the bus runs during the day but not at night, not for its mathematical simplicity but because it matches her understanding of how the world works. We do not contest this point of view. We model simplicity as an a-priori preference order over theories, with the choice of a theory then reflecting this a-priori preference as well as its goodness of fit. Our primary interpretation of this order is that of simplicity, but such an a-priori order over theories may result from other considerations, including a Bayesian prior over theories, aesthetic considerations, notions of plausibility, social norms, computational costs, and so on.[2] Our point is that effective learning requires *some* such information—it cannot be based on evidence alone. Induction is doomed to failure unless coupled with some exogenous prejudice, whatever its origin or interpretation.

There is a vast body of literature in statistics and machine learning that deals with statistical learning. In particular, the Vapnik-Chervonenkis ([17, 18]) theory, recently applied to decision theory by Al-Najjar [2], deals with the rate at which one can learn the probabilities of several events simultaneously. In contrast to this literature, we are interested in optimal learning without assuming that there is an underlying probability law from which the learner can sample in an independent-and-identically-distributed manner. Rather, our main concern is the learning of a pattern that has been selected once and for all at the beginning of time. For example, while statistical learning might be concerned with the prediction of the weather on a given day, assuming that it follows an i.i.d. distribution, our main concern

[2]The key features of the relation are that it is a weak order and that every theory has only finitely many theories that are preferred to it. Any such order can serve as the "simplicity" order for our purposes.

would be in determining whether global warming is underway.

There are many economic problems that require classical statistical learning. However, there are also many problems that do not have sufficiently many repetitions to make the i.i.d. sampling a reasonable assumption. For example, one might reasonably describe candidates for admission to a graduate school as a long sequence of i.i.d. (conditional on observable characteristics) repetitions. By contrast, when deciding whether to get married, one has a limited database about oneself and one's prospective spouse. Similarly, predicting whether a particular customer will make a purchase falls under the classical theory of statistical learning, but predicting whether a stock market crash or a war are in the offing do not. In some variants of our model, we consider a random process that generates i.i.d. "error" terms. However, the focus remains on the basic model selection problem: ascertaining which is the (potentially nonstationary) underlying process. We are thus interested in a learning problem that is non-trivial even for deterministic processes.

1.2 Results

We begin in Section 3 with a simple deterministic model that conveys the basic point. For an infinitely patient reasoner, a preference for simplicity weakly dominates indifference to simplicity, provided that the environment is not a "simplicity trap," i.e., provided that the actual data generating process is not malevolent enough to outsmart a simplicity-preferring reasoner (Section 3.1).[3]

Our result rests on a simple enumeration argument: a simplicity-liking reasoner will eliminate incorrect theories until she gets to the correct one, thereafter predicting the environment with precision (assuming that the environment is not infinitely complex). An agent who relentlessly chases goodness of fit may well never settle on the correct theory, being ultimately doomed to predict no better than chance.

What if the environment is a simplicity trap? Expanding our model to allow such possibilities, we show that a "cautious" preference for simplicity

[3]When the reasoner is not infinitely patient, similar results hold with the appropriate order of quantifiers: for every distribution over data generating processes, sufficiently patient reasoners will be better off preferring simpler theories to more complex ones (Section 3.5). However, given a degree of patience, there could be sufficiently complex environments that the preference for simplicity can be detrimental.

provides protection against simplicity traps and weakly dominates theory selection criteria that do not take simplicity into account (Section 3.3).

Is our reasoner's enumeration task computable? If not, is it reasonable to assume that the reasoner performs it? In response, Section 3.4 provides a computable version of the basic result, requiring both theories and the reasoner's strategy to be implementable by Turing machines. The result invokes a simplicity notion based on a combination of program complexity and computational complexity, and again relies on caution against simplicity traps.

The agents in our examples had to choose among theories that fitted the data perfectly, so that simplicity had no cost. More generally, there will be a trade-off between simplicity and likelihood (or goodness of fit). Section 4 extends the results to more realistic settings in which the world about which the agent reasons is random rather than deterministic. Our result that the agent cannot simply rely on goodness-of-fit comparisons is strengthened in this environment. It is an optimal strategy for the agent to regularly reject theories that provide *superior* fits in favor of less successful but simpler ones, for much the same reasons that statisticians prefer simpler models and scientists prefer more parsimonious theories in order to avoid the dangers of overfitting their data. To ensure this preference for simplicity is successful, however, it must be coupled with a preference for stability. The agent will thus embrace a theory promising enhanced explanatory power only if it is sufficiently simple *and* has provided sufficiently good fit for sufficiently long time.

1.3 Implications

We view this exercise as making the case that inference cannot effectively be based on likelihood arguments alone. Simply observing that one theory fits the data better than another is not sufficient to prefer the former over the latter. Instead, one must also argue that the candidate theory fares well in terms of auxiliary criteria such as simplicity and stability.

Our results also suggest that an agent's performance, as well as our intuitive judgment of his "intelligence," need not be monotone with respect to cognitive abilities. In our examples, a reasoner who is computationally unbounded and has no taste for simplicity fares just as badly as a reasoner who has no computational power whatsoever. Cognitive limitations may be a way to induce a preference for simplicity, thereby improving the performance of

an agent who would otherwise prefer higher likelihood alone.[4] The literature on bounded rationality may thus benefit from a distinction between bounds on rationality that are typically detrimental and those that may be useful. Some are inevitable limitations and some are blessings in disguise. Economic models of idealized rational agents may wish to ignore limitations of the first type, but perhaps not of the second type. The study of the reasoning and behavior of intelligent agents may therefore be enriched by evolutionary psychology and a closer study of the origins of various limitations of the human mind.

2 The Model

2.1 The Environment

We consider a repeated prediction game. In each period, an agent has to predict an observation from the set $\{0, 1\}$. His stage payoff is 1 for a correct prediction and 0 for an incorrect one.

The agent has a history-dependent preference relation over theories in a set T, and at each stage he makes a prediction according to a theory that is a maximizer of this relation for the given history. Each theory is a function from all conceivable histories to predictions.

Before the repeated prediction game begins, two other players are called upon to make their choices in a one-shot meta-game. *Fate* chooses the data generating process the agent will face. Formally, data generating processes, like the agent's theories, are functions from histories to observations or, more generally, to probabilities over observations. The data generating process does not depend on the agent's predictions.

The second player in the meta-game is *Evolution*, who chooses the preference relation of the agent over theories. In the meta-game, Evolution's payoff will be the long-run average payoff of the agent, to be defined shortly.[5]

Two views of the problem we study are available. We can view Evolution as a process by which the agent is programmed with a preference relation over

[4] To paraphrase Voltaire, if computational costs did not exist, it would be necessary to invent them.

[5] Fate's payoff is immaterial as Fate is non-strategic. Fate thus plays a role analogous to "nature" in the conventional description of a decision problem as a game "against nature." Since Evolution can also be thought of as Mother Nature, we decided to avoid the term "nature."

theories. More precisely, we would imagine an evolutionary process in which mutations regularly give rise to agents with various preference relations over theories, with a process of natural selection leading to outcomes in which those preference relations giving rise to relatively high payoffs survive while other preference relations disappear from the population. Alternatively, we can take a normative approach, viewing the agent's preference relation over theories as the result of a decision problem under uncertainty. The states of the world correspond to the various data generating processes that might be chosen by Fate, while Evolution in this case is simply a decision maker charged with choosing a preference relation over theories. In either interpretation, the relevant payoff is a long-run average of the agent's probability of success in predicting the observations of the data generating process.[6]

Observations. At the beginning of each period $n \geq 0$, the agent observes a profile of variables $x_n = (x_n^1, ..., x_n^m) \in \{0,1\}^m \equiv X$. The agent then predicts the value of another variable, $y_n \in \{0,1\}$, to be revealed at the end of period n. We assume the x_n are pre-determined. That is, we fix a sequence $\{x_n\}_{n\geq 0}$ and conduct the discussion relative to this sequence, without specifying the process that generated it.[7]

A history of length $n \geq 0$ is a sequence $h_n = ((x_0, y_0), ..., (x_{n-1}, y_{n-1}), x_n)$. The set of all histories of length n is denoted by $H_n = (X \times \{0,1\})^n \times X$. The set of all histories is $H = \cup_{n\geq 0} H_n$.

Fate. A *data generating process* is a function $d : H \to [0,1]$, with $d(h_n)$ being the probability that $y_n = 1$ given history h_n. The set of all data generating processes is thus $[0,1]^H$. We will typically be interested in problems in which Fate is constrained to choose a data generating process from a certain subset $D \subset [0,1]^H$. For example, $D_0 = \{d \in [0,1]^H |\ d(h) \in \{0,1\}\ \forall h \in H\}$ denotes the set of all deterministic data generating processes.

[6] Our initial assumption that agents are infinitely lived and infinitely patient is relaxed in Section 3.5.

[7] None of our results depends on the characteristics of this data generating process or on realizations of the data having particular properties. In a more general model, some of these variables might be determined by the agent, who might decide to perform experiments and test various theories. Our focus at this point is on learning without experimentation.

The Agent. In each period $n \geq 0$, the agent chooses a theory, denoted by t_n, from a set T. We assume that T is itself a set of data generating processes and typically also assume that $D \subset T$, so that the agent does not face the impossible task of trying to learn a data generating process of which he cannot conceive. We then typically further simplify the agent's learning process by assuming that $D = T$.

The agent uses the theory t_n to predict the period-n value y_n given history h_n. If $t_n(h_n) > 0.5$, then the agent (optimally) predicts $y_n = 1$ with probability 1. He predicts $y_n = 0$ if $t_n(h_n) < 0.5$, and predicts 0 and 1 with equal probability if $t_n(h_n) = 0.5$.

Evolution. Evolution endows the agent with a relation that the agent uses to choose the theory t_n. In particular, for every history $h \in H$, the agent applies a relation $\succsim_h \subset T \times T$ to the set T of theories. We assume that, for every h, $\succsim_h$ is complete and transitive, and that it has maximal elements. We define

$$B_{\succsim_h} = \{t \in T \,|\, t \succsim_h t' \quad \forall t' \in T\}$$

to be the set of "best" theories in the eyes of the agent faced with history h and characterized by $\succsim_h$. The agent's choice following history h is unambiguous if $B_{\succsim_h}$ is a singleton, but this may often fail to be the case. What does the agent do if there are a number of theories in $B_{\succsim_h}$? Our basic assumption is that the agent in this situation treats the various best theories symmetrically, in the sense that he makes a choice that wherever possible exhibits no bias for theories that predict 0 versus theories that predict 1 in the next observation.

Assumption 1 *The agent chooses from $B_{\succsim_h}$ according to a measure $\mu_{B_{\succsim_h}}$ on $B_{\succsim_h}$ satisfying*

$$\mu_{B_{\succsim_h}}\left(\left\{t \in B_{\succsim_h} \,|\, t(h) < 0.5\right\}\right) = \mu_{B_{\succsim_h}}\left(\left\{t \in B_{\succsim_h} \,|\, t(h) > 0.5\right\}\right)$$

whenever

$$\left\{t \in B_{\succsim_h} \,|\, t(h) < 0.5\right\}, \left\{t \in B_{\succsim_h} \,|\, t(h) > 0.5\right\} \neq \varnothing.$$

We explain in Section 3.1 why this assumption, which is not needed for all of our results, is especially natural in the settings we examine.

One might protest that if an agent has observed a relentless stream of 0s, perhaps many thousands long, then it surely makes no sense to view a 0 and a 1 as equally likely as the next observation. We agree: this is precisely the reasoning behind the idea that simplicity may be a valuable aid to making decisions. However, we view the relation $\succsim$ (rather than the choice from $B_{\succsim_h}$) as capturing this type of reasoning. After observing a history h consisting entirely of 0s, for example, the relation $\succsim_h$ is likely to give pride of place to theories predicting a zero on the next round. Assumption 1 applies only when the agent has exhausted all concerns about goodness of fit, simplicity, plausibility, and so on, and still entertains theories predicting a 0 and theories predicting a 1 in the next period. Only then does Assumption 1 require symmetric treatment of such theories.

Payoffs. Let $d(h_n) \in [0,1]$ and $\hat{t}_n(h_n) \in \{0,1\}$ be the outcome of the data generating process d and the agent's prediction, respectively, given history h_n. The payoff to the agent in period n is defined by the probability of guessing y_n correctly, that is,

$$p(d, \hat{t}_n, h_n) = d(h_n)\hat{t}_n(h_n) + (1 - d(h_n))\left(1 - \hat{t}_n(h)\right).$$

Intuitively, we would like to take the long-term payoff to the agent to be the expected value of

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T-1} p(d, \hat{t}_n, h_n), \tag{2}$$

where the expectation captures the potential randomness in the outcomes produced by the data generating process, the agent's choice of theories, and the resulting predictions. However, this limit need not exist. We let

$$\Lambda(\{p(d, \hat{t}_n, h_n)\}_{n=0}^{\infty})$$

be a Banach limit defined on the set of sequences $\{p(d, \hat{t}_n, h_n)\}_{n=0}^{\infty}$, and let the agent's payoff $P(d, \succsim)$, when facing data generating process d and using relation $\succsim = \{\succsim_h\}_{h\in H}$ to choose theories, be given by the expected value of Λ. The key property of Banach limits we need is that

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T-1} p(d, \hat{t}_n, h_n) \leq \Lambda(\{p(d, \hat{t}_n, h_n)\}_{n=0}^{\infty}) \leq \limsup_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T-1} p(d, \hat{t}_n, h_n).$$

2.2 Choosing Theories

Our main interest is in the agent's relation $\succsim$ for choosing theories. We introduce here our two primary contenders.

2.2.1 The Likelihood Relation

We can view Lisa and Lloyd in our introductory examples as employing the *likelihood relation*. The likelihood relation chooses theories based solely on the evidence, selecting theories that fit the data best. Formally, for $h_n \in H_n$, let

$$L(t, h_n) = \Pi_{j=0}^{n-1}[t(h_j)y_j + (1 - t(h_j))(1 - y_j)].$$

This is the likelihood of theory t given history h_n. The likelihood relation $\succsim^L$ ranks theories after any history h by their likelihood:

$$\forall h \in H, \quad t \succsim_h^L t' \quad \iff \quad L(t, h) \geq L(t', h).$$

The likelihood relation thus calls for agents to base their inferences on their data, and on no other criterion.

In the simple case where $T = D_0$, i.e., when only deterministic theories are considered, $\succsim_h^L$ boils down to two equivalence classes. All theories that perfectly fit the data are equivalent, having $L(t, h) = 1$, and they are all preferred to all theories that have been refuted by the data, where the latter are also equivalent to each other and satisfy $L(t, h) = 0$.

2.2.2 The Simplicity Relation

Sally and Sam in our examples have a taste for simplicity. The simplicity relation takes complexity into account, though only as a secondary criterion. To define this theory-selection procedure, we first order T according to the simplicity of its elements. An order $\succcurlyeq^S \subset T \times T$ is a *simplicity order* if:[8]

$$\succcurlyeq^S \text{ is a weak order (i.e., complete and transitive)} \tag{3}$$

[8]In an effort to keep things straight, we use $\succsim$ to denote a relation by which the agent chooses theories, and $\succcurlyeq$ to denote a simplicity order over theories. We similarly associate the label "relation" with the former and "order" with the latter (though they have the same properties, i.e., each is complete and transitive). Observe that $\succcurlyeq$ is defined a-priori, independently of history, whereas $\succsim$ is a function of $h \in H$.

and

$$T' \subset T \text{ countable } \Longrightarrow \#\left\{t' \in T' \,|\, t' \succcurlyeq^S t\right\} < \infty \qquad \forall t \in T'. \tag{4}$$

If $\succcurlyeq^S$ is a simplicity order on T and $T' \subset T$ is countable, then there exist $\succcurlyeq^S$-maximal elements in T'.

In the absence of condition (4), the definition of a simplicity order would be consistent with the trivial order $\succcurlyeq^S = T \times T$, according to which no theory is strictly simpler than another. More generally, we would like to rule out the case that the simplicity order is permissive enough to allow for infinitely many strategies to be equally simple.

If the set of data generating processes is countable, one way to ensure that (3) and (4) hold is to enumerate T and set $t_i \succ^S t_{i+1}$ for every $i \geq 1$, so that $T = \{t_1, t_2, ...\}$ is ordered according to increasing complexity. Our definition is less demanding, and allows for non-singleton equivalence classes of the order $\sim^S$, but not for infinite ones. Nonetheless, under the assumption that T is countable, simplicity orders are closely related to enumerations of T. Specifically, for every simplicity order $\succcurlyeq^S$ there exists an enumeration $T = \{t_1, t_2, ...\}$ such that $t_i \succcurlyeq^S t_{i+1}$, with strict preference $\succ^S$ occurring for infinitely many i's.[9] Alternatively, $\succcurlyeq^S$ is a simplicity order if and only if it can be represented by a complexity function $C : T \to \mathbb{N}$ such that

$$t \succcurlyeq^S t' \quad \Longleftrightarrow \quad C(t) \leq C(t') \tag{5}$$

and

$$|C^{-1}(k)| < \infty \qquad \forall k \in \mathbb{N}.$$

Given a simplicity order $\succcurlyeq^S$, we define the associated simplicity relation $\succsim^{LS}$ for choosing theories as follows:

$$\forall h \in H, \quad t \succsim_h^{LS} t' \iff \begin{cases} & \{t \succ_h^L t'\} \\ or & \{t \sim_h^L t' \ \ and \ \ t \succcurlyeq^S t'\} \end{cases}.$$

The relation $\succsim^{LS}$ thus uses the simplicity order $\succcurlyeq^S$ to choose among those theories with the highest likelihood. The likelihood and simplicity relations

[9] While there are obviously many different enumerations of T, and hence many complexity functions C with their induced simplicity orders $\succcurlyeq^S$, they cannot be too different in the following sense. Let C_1 and C_2 be two complexity functions. Then, for every k there exists $l = l(k)$ such that $C_1(t) > l$ implies $C_2(t) > k$. That is, a theory that is sufficiently complex with respect to C_1 will also be complex with respect to C_2.

$\succsim^L$ and $\succsim^{LS}$ agree in that they only choose theories with maximal likelihoods, with the likelihood relation being indifferent over such theories and the simplicity relation choosing the simplest one.

3 Preference for Simplicity

This section derives a preference for simplicity in an elementary deterministic model. The key simplification is contained in the following assumption, which puts some structure on the data generating processes. Its first two parts require the set of data generating processes to be simple enough to be learned, and the final part requires it be rich enough to describe any possible finite sequence of observations. The latter is intended to rule out trivial cases in which a finite set of observations suffices to single out a unique theory, i.e., cases where the problem of induction does not arise.

Assumption 2

[2.1] $D \subset D_0$.

[2.2] $D = T$ *is countable.*

[2.3] *For every history* $h \in H$ *there exists* $d \in D$ *such that* $L(d, h) = 1$.

Given Assumption 2.1, the remaining requirements are satisfied if $D = T$ is the set D_0^H of all Turing machines generating functions $d \in D_0$ (i.e. Turing machines that accept elements of the set H as inputs, halt, and produce outputs from the set $\{0, 1\}$). The countability restriction will be discussed and relaxed in Section 3.3 below. We view the restriction to deterministic data generating processes as the substantive assumption here. One natural simplicity relation $\succsim^S$ on computable processes is induced by Kolmogorov's complexity measure, namely, the minimal description length of a program that generates the process in a given computer language.

The set D_0^H of all Turing machines generating functions $d \in D_0$ has the property that for any history of observations h, and for every data generating process d in D_0^H consistent with h, there is another data generating process in D_0^H that is also consistent with h but whose subsequent datum will be precisely the opposite of d, generating a 0 whenever d produces a 1 and vice versa. As a result, the sets $\left\{t \in B_{\succsim^L_h} \,|\, t(h) = 0\right\}$ and $\left\{t \in B_{\succsim^L_h} \,|\, t(h) = 1\right\}$ will not only be non-empty but will be symmetric in their treatment of the next observation. Assumption 1, requiring that an unbiased choice be made from these sets under the likelihood relation, is then quite natural.

D	Set of data generating processes $\left(\subset [0,1]^H\right)$.
D_0	Set of deterministic data generating processes (i.e., with outputs $\{0,1\}$).
D_0^T	Set of Turing machines with inputs H and outputs $\{0,1\}$.
D_0^H	Set of Turing machines in D_0^T that halt for all $h \in H$.
D_0^B	Set of Turing machines in D_0^T with bounded halting time.
D_ε	Set of data generating processes with outputs $\{\varepsilon, 1-\varepsilon\}$.

Figure 1: Data Generating Processes. In each case, "Set of Turing machines...." should be read "set of data generating process that can be implemented by a Turing machine...."

In the course of our discussion, we will consider several possibilities for the set D. It is useful for future reference to collect the notation for these various sets in Figure 1.

3.1 Comparison of Payoffs

Recall that $P(d, \succsim)$ is the payoff function defined by a Banach limit corresponding to (2), given that Fate has chosen data generating process d and Evolution has endowed the agent with relation $\succsim$.

Proposition 1 *Let Assumption 2 hold.*

[1.1] For every simplicity order $\succcurlyeq^S$ and every $d \in D$, $P(d, \succsim^{LS}) = 1$.

[1.2] If Assumption 1 also holds, then $P(d, \succsim^L) = .5$, and hence for every simplicity order $\succcurlyeq^S$, $\succsim^{LS}$ strictly dominates $\succsim^L$.

Proposition 1.2 holds for much the same reason that statisticians and machine learning theorists are concerned with "overfitting," even though there are no statistical errors in our simple model and hence the term overfitting does not quite apply.[10] The agent has to choose among theories that have

[10]The preference for simplicity among theories that match the data equally well may be viewed as a tendency not to "overkill," rather than not to overfit. Overfitting refers to a willingness to complicate the theory for relatively small improvement in the goodness of fit, while overkilling might be thought of as the willingness to complicate the theory for no improvement in the goodness of fit whatsoever. Section 4 presents a more general model, in which statistical errors are allowed, to introduce a non-trivial trade-off between the simplicity of a theory and its accuracy. The finding that preference for simplicity is evolutionarily selected is then more closely related to the dangers of overfitting in the standard statistical sense.

been refuted and theories that match the data perfectly. Realizing that one can always find many theories that match the data perfectly, with a variety of differing predictions, standard statistics leads us to prefer theories that are simpler than the data that they fit. The proof of Proposition 1.2 similarly relies on the fact that, in the absence of preference for simplicity, one is at a loss in selecting a theory. Many theories match the data perfectly, but their predictions vary. There is nothing to ensure that we ever prefer the correct theory to other theories, and our prediction ends up being completely random.

It is important to observe that our preference for simplicity is an interpretation of an abstract order $\succcurlyeq^S$. In many cases, this order will correspond to an intuitive sense of simplicity. However, Proposition 1 does not assume any particular structure on the notion of simplicity and no particular measure of simplicity can be singled out as the right one. Any consideration that allows one to rank theories by an order satisfying (3)–(4) would result in a simplicity relation $\succsim^{LS}$ that dominates $\succsim^{L}$.

3.2 Other Contenders

The likelihood relation and the simplicity relation by no means exhaust all strategies one can imagine for choosing theories. We devote this section to the discussion of alternatives that help us understand the driving force behind the success of simplicity.

3.2.1 Ideological

"Ideological" strategies ignore evidence completely and relentlessly choose a single element $t \in T$. Clearly, for every data generating process $d \in D$ there exists an ideological strategy that chooses d and hence gives payoff 1. Why doesn't Evolution simply give the agent this preference relation—essentially, explain the world to the agent—and be done with it? The only means evolution has for choosing the right data generating process is trial-and-error. Our view is that the data generating process d is a characteristic of the environment that changes too quickly for trial-and-error to keep up.[11]

[11] "Too quickly" is a relative term. It may well be that the data generating process has been fixed since the start of human history, and yet Evolution had time to go through so little of the countable set T as to have no hope of hitting the true data generating process. We make the idea of "too quickly" operational in our model by thinking of every agent as

From a normative or statistical interpretation, the question simply does not arise: there are many ideological strategies, and finding the right one is precisely the problem of inductive inference.

3.2.2 Bayesian

Suppose that the data generating process d facing the agent is chosen according to a probability measure λ on D. Then the agent's payoff is maximized by a strategy that predicts 1 in period n after history h_n if and only if

$$\lambda\{d : L(d, h_n) = 1, d(n) = 1\} > \lambda\{d : L(d, h_n) = 1, d(n) = 0\},$$

that is, there is no way to do better than to make the agent Bayesian. Why doesn't our inquiry end with this observation?

Our approach is compatible with Bayesianism. Specifically, if the set of conceivable theories T is countable (cf. Assumption 2) and the agent has a Bayesian prior over T, then the relation "has at least as high a prior as" can be viewed as a simplicity relation—it is a weak order that is monotonically decreasing along an enumeration of the theories, with finite equivalence classes. In other words, a Bayesian prior defines a simplicity relation. Conversely, one may use a simplicity relation to define a Bayesian prior: simpler theories are considered more likely.

There are a continuum of priors that are consistent with a given simplicity relation. These priors are all equivalent in our model, because we suggest that the agent choose a most-likely theory to generate the next prediction—a practice we dub "simplicism." By contrast, the Bayesian approach constructs an expected prediction, using all possible theories. However, after a long enough time, the Bayesian will have a posterior probability close to 1 for the correct theory, and (assuming its prediction is deterministic) will also converge to making predictions as if he used simplicism.

Simplicism and Bayesianism are thus similar both in their eventual selection of the "best" theory and in the predictions they eventually generate. Yet, simplicism is computationally and cognitively simpler than Bayesianism in several important ways.[12] First, in order to implement simplicism, one need not conceive of all possible theories. For example, a scientist who finds

receiving an idiosyncratic draw from D.

[12]Clearly, simplicism is a more faithful model of how scientists actually think. For example, relativity theory did not wait to be discovered by Einstein (or Poincare) because until his time its posterior probability was too low to be prominent. It was simply not

the simplest theory that explains the data may go on generating predictions even if he has not even imagined alternative explanations. Second, in order to generate predictions, simplicism requires only an ordering over the theories, rather than a quantification thereof. It is an easier task to ask which theory is simpler than to assign a numerical value to the a-priori likelihood of a theory. Relatedly, a Bayesian approach requires that an "at least as likely as" relation be defined for every two subsets of theories, and not just for every pair of theories. Therefore, even if we restrict attention to qualitative comparisons, a Bayesian needs to rank all elements in 2^T, rather than elements in T itself.

In summary, our justification of a preference for simplicity is compatible with a justification of a Bayesian approach. Both approaches use subjective notions—the preference for simplicity or the prior—that need not be "correct" to be useful in the inductive inference task. We continue to discuss "simplicity" as our preferred interpretation of the a-priori preference between theories, thinking that bounded rationality would tip the scales in this direction (see Section 5.2), but would not object to others adopting a Bayesian interpretation.

3.2.3 Inertial

The simplicity relation has two components. It ensures that the agent will not abandon a theory that fits the data perfectly well. In addition, it essentially enumerates the theories, ensuring that an agent will eventually "try" all of them, if needed. One might suspect that it would be sufficient to impose the first requirement by assuming that agents do not abandon a theory until receiving evidence of its falsity. In particular, the proof of Proposition 1 shows that an agent guided by the likelihood relation falters because every period there is a multitude of theories with perfect likelihood scores, including the truth and a host of imposters. The agent's arbitrary choice from this set implies that even if he hits upon the truth, he soon abandons it in favor of another seemingly equivalent theory. Will it not suffice to assume that the agent sticks to something that has worked in the past?

The phenomenon of inertia, or a preference for a status quo, is familiar from casual observations as well as from psychological studies. Indeed, Kuhn

conceived of by his predecessors. Indeed, if scientists were Bayesian, and were only updating prior to posterior probabilities based on data, there would be little point in hiring theoretical scientists in the first place.

[11], suggested that scientists tend to cling to old theories rather than adopt those theories that fit the data best.

To see if inertia suffices for effective learning, we define the *inertial* relation as that selecting the theory chosen in the previous period if the latter maximizes the likelihood function, and otherwise choosing as does the likelihood relation. Formally, define $\succsim^{LI}$ as follows for all $n > 1$,

$$\forall h \in H, \qquad t \succsim^{LI}_{h} t' \iff \begin{cases} & \{L(t,h) > L(t',h)\} \\ or & \{L(t,h) = L(t',h) \ \ and \ \ t = t_{n-1}\} \\ or & \{L(t,h) = L(t',h) \ \ and \ \ t, t' \neq t_{n-1}\} \end{cases},$$

with $t \sim^{LI}_{h_0} t'$ for all t, t', so that in the absence of any evidence, all theories are equally likely.

The following example shows that inertia alone does not suffice to ensure effective learning.

Example 3: Let (for this example only) D consist of the following set of deterministic theories:

$$\left\{ y \in \{0,1\}^{\mathbb{N}} \mid \exists n \geq 0, \ \ y(k) = 0 \ \forall k \geq n \right\}.$$

The possible data generating processes are thus all those that generate only 0 from some point on. For example, the theories may be describing the availability of a random resource, which is known to be depletable, but whose date of ultimate exhaustion is uncertain.

For every h_n, let the selection rule over the infinite set $B_{\succsim^{LI}_{h_n}}$ be given by

$$\mu_{B_{\succsim^{LI}_{h_n}}}\left(t^{n+k}\right) = \frac{1}{2^{k+1}} \qquad k = 0, \ldots, \tag{6}$$

where, for all histories and all k,

$$t^{n+k}(h_{n+k}) = 1; \quad t^{n+k}(h_l) = 0 \quad \forall l \neq n+k. \tag{7}$$

Under this selection rule, theories predicting a 0 on the next step are equally likely as theories predicting a 1, in accordance with (1). Assume Fate has chosen the data generating process according to which $y_n = 0$ for all n. Consider $\succsim^{LI}_{h_n}$ for a history h_n consisting of n 0s. Then given (6)–(7), $\succsim^{LI}$ will choose a theory whose first 1 appears according to a geometric distribution

with parameter 0.5. The expected number of periods for which this theory will match the observations is

$$\sum_{i=0}^{\infty} \frac{i}{2^{i+1}} = 1.$$

It is then a straightforward calculation that $P(y_0, \succsim^{LI}) = 0.5$. ■

The difficulty in this example is that the selection rule over the various sets $B_{\succsim^{LI}_{h_n}}$ routinely ignores the correct theory. This suggests that inertial relations will ensure effective learning only if they are coupled with a selection rule that is sufficiently likely to select the correct theory. Proposition (2) shows that inertia can then have evolutionary value, in effect serving as a safeguard against the excessive fickleness of random choice.

Assumption 3 *There exists a strictly positive measure λ on the countable set D such that for any $h \in H$, $\mu_{B_{\succsim_h}}$ equals λ conditioned on $B_{\succsim_h}$.*

Proposition 2 *Under Assumptions 2 and 3, for all $d \in D$, $P(d, \succsim^{LI}_h) = 1$.*

Behind this result lies the observation that a sufficiently stationary theory selection process is guaranteed to select the correct theory, d, at least once. Once d has been chosen, inertia ensures that it will not be abandoned, and hence the optimal payoff is obtained.

3.2.4 Exploitation and Exploration

The simplicity relation and the inertial relation can both be viewed as special cases of the principle of "exploitation and exploration." The agent exploits theories that have worked by sticking with them, while effectively exploring new theories when necessary. In the case of simplicity, the stability of a given simplicity ordering ensures that a theory that fits the data is not abandoned, while the simplicity enumeration allows the agent to "try out" all theories (as long as a perfect fit has not been found). The likelihood relation's lack of the first characteristic dooms its adherents to randomness. The inertial relation matches the performance of the simplicity relation in terms of exploitation, but requires an additional assumption on the strategy selection process to provide effective exploration.

There may be many other strategies that guarantee exploitation and exploration, though our notion of simplicity orders is general enough to cover all strategies that are based on an enumeration of the theories. We focus on simplicity and inertial relations partly because they appear to be intuitive and easily implementable.

3.3 Countability

We have assumed that $D = T$ is countable. The countability of T may seem quite restrictive. Indeed, most statistical models allow continuous parameters, and thereby seemingly refer to uncountable families of processes. However, our inclination is to be persuaded by Church's thesis—if the agent can make a particular set of predictions, then there must be a Turing machine generating these predictions (Hopcraft and Ullman [7, Chapter 7]), and hence the set T can reasonably be taken to be countable.[13]

But this limitation on the agent's cognitive abilities need not be shared by Fate. We may well have a set D that is an uncountable superset of T. How will a simplicity seeking agent fare then? Worse still, what if Fate is malevolent, using a (noncomputable) strategy that predicts the agent's (computable) predictions in order to then generate unpredicted observations? To investigate this possibility, we retain the assumption that $T \subset D_0$ is countable, but allow $D \subset D_0$ to be a superset of T.

The standard way for the agent to protect himself against a malevolent Fate is to randomize. Specifically, for a simplicity order $\succcurlyeq^S$ and for $\varepsilon > 0$, let the relation $\succsim^{LS,\varepsilon}$ be defined by augmenting $\succsim^{LS}$ with a "safety net." If the average payoff at history h_n is lower than $0.5 - \varepsilon/\log n$, then $\succsim^{LS,\varepsilon}_{h_n} = T \times T$. Otherwise, $\succsim^{LS,\varepsilon}_{h_n} = \succsim^{LS}_{h_n}$.

[13] Alternatively, one may arrive at countability via a more lenient model, in which a Turing machine (or, equivalently, a PASCAL program) can also perform algebraic operations on arbitrary real-valued variables, where the actual computations of these operations are performed by an "oracle" that is not part of the machine's computation. A stricter interpretation of computability, which does not resort to "oracles," would restrict attention to statistical models in which all parameters are computable numbers. A number $x \in \mathbb{R}$ is *computable* if there exists a Turing machine M that, given the description of any rational $\varepsilon > 0$, performs a computation that halts, and writes a number $M(\varepsilon) \in \mathbb{Q}$ such that $|M(\varepsilon) - x| < \varepsilon$. All rational numbers are computable, but so is any irrational number that can be described by a well-defined algorithm, including algebraic irrational numbers (such as $\sqrt{2}$), e, and π.

Proposition 3 *Let $T \subset D_0$ be countable. Under Assumptions 1, 2.1 and 2.3 (but allowing $D \subset D_0$ to be a superset of T), $\succsim^{LS,\varepsilon}$ weakly dominates $\succsim^{L}$ for every simplicity relation $\succcurlyeq^{S}$, with $\succsim^{LS,\varepsilon}$ performing strictly better for data generating processes $d \in T$.*

Proposition 3 can be interpreted as saying that a preference for simplicity could be evolutionarily favored even when a cruel Fate ensnares the agent to believe the world is simple, only to prove her wrong.

3.4 Computability

We have justified the assumption that T is countable by appealing to computability arguments, in the form of an assumption that the agent can only implement predictions generated by a Turing machine. Continuing in this spirit, we now take computability issues more seriously. Let us first restrict Fate to the set D_0^H of deterministic data generating processes implementable by Turing machines that halt after every input $h \in H$.

In contrast, we now allow the agent to consider the set D_0^T of all Turing machines, even those that do not always halt. It is a relatively easy task for the agent to enumerate all Turing machines, but it is not an easy task to check which of them do indeed define a data generating process.[14] A model that respects the agents' computability constraints must then allow the set T to include *pseudo-theories*: all machines that can be written in a certain language (and therefore appear to define a data generating process), even if they may not halt for all histories. Clearly, this additional freedom cannot help the agent: if, at a given history h, the agent chooses a machine that does not halt for that history, he will never be able to make a prediction (in which case we take his payoff to be 0). However, "helping" the agent by assuming that $T \subset D_0^H$ would be unreasonable, as it would be tantamount to magically endowing the agent with the ability to solve the celebrated halting problem.[15]

[14] One could simulate the computation of any given machine given input h, but there is no way to distinguish between computations that take a long time and computations that never end.

[15] Formally speaking, the objects of choice for the agent are not theories but descriptions thereof. A rigorous treatment of this problem would call for the definition of a formal language and of a means of describing programs in that language. Some descriptions give rise to well-defined theories (i.e., that halt for every history), whereas others would not.

We also restrict the agent to relations $\succsim$ that are computable, in the sense that for every $h \in H$, the choice made by the relation $\succsim_h$ from the set $B_{\succsim_h} \subset D_0^T$ could itself be implemented by a Turing machine that inevitably halts. This restriction is a binding constraint for some data generating processes:

Proposition 4 *For every computable relation $\succsim \subset D_0^T \times D_0^T$, there exists a data generating process $d \in D_0^H$ such that $P(d, \succsim) \leq 0.5$.*

Proposition 4 imposes a bound on what can be guaranteed by a computable strategy, in the sense that any such strategy must fare no better than chance against some data generating processes. The proof consists of observing that if the agent's strategy is computable, then one may always construct a malevolent strategy d that mimics the agent's computation and chooses an observation that refutes it.

The malevolent strategy d used to prove Proposition 4 is quite far from most statistical models. Fate might be quite complex, but it is hard to imagine that Fate spitefully lures the agent into believing in a simple environment, only to refute this belief period after period. Will a more neutral model of Fate allow a possibility result? One way to obtain a more realistic set of data generating processes is to limit their computation time. Specifically, let D_0^B be the set of data generating processes that are implementable by Turing machines that halt within a bounded number of steps. That is, for $d \in D_0^B$ there exists a Turing machine $M(d)$ and an integer $K(d)$ such that, for every history h_n and attendant prediction y_n, the computation of $M(d)$ on h_n halts within $K(d)$ steps, producing y_n.

The agent is restricted to have a simplicity order that is represented by a computable function $C : D_0^T \to \mathbb{N}$, so that

$$C(t) \leq C(t') \qquad \Longleftrightarrow \qquad t \succcurlyeq^S t'.$$

Thus, because C is computable, the agent can compute $\succcurlyeq^S$.

The following result adapts simplicity-based rankings to the computable set-up.

Proposition 5 *For every computable simplicity order $\succcurlyeq^S \subset D_0^T \times D_0^T$, there exists a computable relation $\succsim$ with each $\succsim_h \subset D_0^T \times D_0^T$ such that*

In such a model, every theory would have infinitely many equivalent descriptions. Thus, the function that maps descriptions to theories is not defined for all descriptions and is not one-to-one. To simplify notation, we do not pursue these formal details.

(5.1) $P(\succsim, d) = 1$ *for every* $d \in D_0^B$*;*

(5.2) for every $d, d' \in D_0^B$ *there exists* N *such that, for every* $n \geq N$ *and every* $h \in H_n$ *for which* $L(d,h) = L(d',h)$ *,*

$$d \succ_h^S d' \Rightarrow d \succ_h d'.$$

Proposition 5 ensures the existence of a computable strategy yielding optimal payoffs, as well as its asymptotic agreement with the (strict part of) the given simplicity ordering $\succcurlyeq^S$ over D_0^T.[16] The relation $\succsim$ cannot follow $\succsim^{LS}$ precisely, but it does so for long enough histories. In other words, it is possible that for a short history the relation $\succsim$ will not reflect the simplicity ranking $\succcurlyeq^S$, but in the long run, any two theories that are equally accurate but not equally simple will be ranked according to $\succcurlyeq^S$.

Observe that most deterministic statistical models encountered in the social sciences are in D_0^B. The deterministic version of models such as linear regression, non-linear regression, as well as many models in machine learning, can be described by an algorithmic rule whose computation time does not depend on the input. A notable exception are time series in economics, where the model describes the dependence of y_n on $\{y_i\}_{i<n}$, and thus the length of the computation increases with the length of history, n.

3.5 Impatient Agent

Suppose that the agent has a discounted payoff criterion,

$$P^\delta(d, \succsim) = (1-\delta) \sum_{n=0}^{\infty} \delta^n p(d, \hat{t}_n, h_n). \tag{8}$$

We assume that Fate chooses a data generating process (as usual, independently across agents) according to a probability measure λ on D.

Simplicity has an advantage provided that the agent is sufficiently patient. In particular, the (omitted) proof of the following is a straightforward modification of the arguments used to prove Proposition 1.2:

[16] It would be natural to think of d as simpler than d' if the Kolmogorov complexity of d is lower than that of d', i.e., if d has a shorter minimal description length than d'. This still leaves some freedom in defining $\succcurlyeq^S$. For instance, one may choose a description in a given programming language, such as PASCAL, as opposed to Turing machines, and one may take the description of constant values into account in the measurement of the description length, or decide to ignore them, and so on.

Proposition 6 *Let Assumptions 1 and 2 hold and let payoffs be given by (8).*

[6.1] For every $d \in D$, there is a discount factor δ^* such that for all $\delta \geq \delta*$,

$$P^\delta(d, \succsim^{LS}) > P^\delta(d, \succsim^{L}).$$

[6.2] If the data generating process is chosen according to density λ on D, then there is δ^* such that for all $\delta > \delta^*$,

$$\int_D P^\delta(d, \succsim^{LS}) d\lambda > \int_D P^\delta(d, \succsim^{L}) d\lambda.$$

4 A Random World

The assumption that the data generating process is deterministic (i.e., that $d(h) \in \{0, 1\}$ for all h) is unrealistic. Worse still, it beclouds the interesting trade-off between likelihood and simplicity in the choice of theories. So far, the choice of simple or random theories was made among the theories that fit the data perfectly—the preference for simplicity involved no cost. A more interesting problem arises when random data generating processes are introduced. In this case, simplicity is no longer a free good, but has a likelihood price tag attached to it. Should the agent be willing to give up a better fit for a simpler theory, and if so, to what extent?

4.1 Uniform Errors

To get some insight into this problem, we begin with a minimal modification of our benchmark model. Define, for $\varepsilon \in (0, 1/2)$,

$$D_\varepsilon = \{d \in [0,1]^H \,|\, d(h) \in \{\varepsilon, 1-\varepsilon\} \ \forall h \in H\}.$$

Thus, D_ε can be thought of as the deterministic data generating processes, D_0, with an error probability of ε added to the output.

The likelihood function, for a theory $t \in D_\varepsilon$ and a history $h \in H_n$, is

$$L(t, h_n) = \Pi_{j=0}^{n-1}(t(h_j)y_j + (1 - t(h_j))(1 - y_j)).$$

In the presence of randomness, the likelihood function will inevitably converge to zero for any theory—its largest possible value in period n is $(1-\varepsilon)^n$,

$$\begin{array}{rcrl} \log(1-\varepsilon) & = & \theta(1) & \text{Maximum possible limiting value.} \\ (1-\varepsilon)\log(1-\varepsilon)+\varepsilon\log\varepsilon & = & \theta(1-\varepsilon) & \text{Value achieved by the data generating process.} \\ \frac{1}{2}\log(1-\varepsilon)+\frac{1}{2}\log\varepsilon & = & \theta\left(\frac{1}{2}\right) & \text{Value achieved by random choice.} \end{array}$$

Figure 2: Key values of the limiting average-log-likelihood function (9).

since the best any theory can do is attach probability $1-\varepsilon$ in each period to the outcome that happens to be realized in that period. This convergence makes the likelihood an awkward standard for comparing theories. It is more convenient to consider the average of the logarithm of the likelihood function,

$$\begin{aligned} l\left(t,h_n\right) &= \frac{1}{n}\log(L(t,h_n)) \\ &= \frac{1}{n}\sum_{j=0}^{n-1}\log\left[t(h_j)y_j+(1-t(h_j))(1-y_j)\right], \end{aligned} \tag{9}$$

which does not converge to zero. We hereafter use "likelihood" to denote the average log likelihood, given by (9).

Let us say that a theory is "correct" in period t if it predicts a 1 with probability $1-\varepsilon$ and a 1 occurs, or if it predicts a 0 with probability $1-\varepsilon$ and a 0 occurs. It is helpful to define the function

$$\theta(p)=p\log(1-\varepsilon)+(1-p)\log\varepsilon.$$

Then $\theta(p)$ is the (average log) likelihood of a theory that has been correct proportion p of the time.

A theory that is correct in every period would give likelihood $\theta(1)$. This is the highest possible likelihood. The theory that corresponds to the data generating process gives a limiting likelihood of $\theta(1-\varepsilon)$, and an agent who always uses the data generating process to predict would achieve payoff $1-\varepsilon$.[17] Predicting randomly would give likelihood $\theta\left(\frac{1}{2}\right)$ and payoff $\frac{1}{2}$. Figure 2 summarizes these observations.

The counterpart of Assumption 2 is now:

Assumption 4

[4.1] $D\subset D_\varepsilon$.

[4.2] $D=T$ *is countable.*

[4.3] For every history $h\in H$ *there exists* $d\in D$ *such that* $l(d,h)=\theta(1)$.

[17] For large n, the likelihood will be approximately$(1-\varepsilon)^{(1-\varepsilon)n}\varepsilon^{\varepsilon n}$ and the average log likelihood $l(d,h)$ will converge to $\theta(1-\varepsilon)$.

Assumption 4.3 indicates that for any finite stream of data, there is a theory that would have been correct in every period. Ex post, one can rationalize anything.

4.2 Tolerance in Learning

The agent could once again adopt a relation over theories that first restricts attention to likelihood-maximizing theories, such as the likelihood relation $\succsim^L$ of Section 2.2.1 or the simplicity relation $\succsim^{LS}$ of Section 2.2.2. In the random environment, this ensures that the agent will eventually *exclude* the data generating process as a possible theory. In each period, the realization may differ from the true theory's prediction with probability ε. Hence, the true theory will eventually almost surely have a likelihood value lower than $\theta(1)$, whereas there will always be other theories with a likelihood value of $\theta(1)$. That is, insisting on maximum-likelihood theories will lead to constant theory hopping.

This suggests that the agent's learning might be more effective if it incorporates some tolerance for inaccuracy. For any $\gamma \in [0,1]$, we say that a theory t is a "γ-best fit" to the data after history h if

$$l\,(t,h) \geq \theta(\gamma).$$

The counterpart of the likelihood relation is then

$$\forall h \in H, \quad t \succsim_h^{L,\gamma} t' \quad \Longleftrightarrow \quad L^\gamma(t,h) \geq L^\gamma(t',h)$$

where

$$L^\gamma(t,h) = \min\{L(t,h), \theta(\gamma)\}.$$

When working with D_0, the likelihood relation $\succsim^L$ separated theories into two classes, those that predicted perfectly and those that did not. Here we again divide theories into two classes, those achieving a likelihood of at least $\theta(\gamma)$ and those that fall short of this goal.

What would be a good value of γ? One suspects that we should set $\gamma < 1-\varepsilon$, since any value $\gamma > 1-\varepsilon$ will eventually surely exclude the true data generating process. However, simply relaxing the likelihood threshold to $\gamma < 1-\varepsilon$ does not suffice if one insists on using the likelihood criterion alone to choose theories. The true theory (if such exists) will not be ruled out, but there is no guarantee that it be selected. An argument analogous to that establishing Proposition 1.2 immediately provides the (omitted) proof of:

Proposition 7 *Let Assumptions 1 and 4 hold. Then* $P(d, \succsim^{L,\gamma}) = \frac{1}{2}$.

Intuitively, whatever the value of γ, the agent has a wealth of theories with likelihoods exceeding $\theta(\gamma)$ from which to choose. In the absence of another selection criterion, the agent is doomed to random prediction.

Once the agent is willing to pay the price of less than maximum likelihood, he can afford to use the additional criterion of simplicity in a meaningful way. Define

$$\forall h \in H, \quad t \succsim_h^{LS,\gamma} t' \iff \begin{cases} \{t \succ_h^{L,\gamma} t'\} \\ \text{or} \quad \{t \sim_h^{L,\gamma} t' \text{ and } t \succcurlyeq^S t'\} \end{cases}.$$

The agent thus chooses the simplest among the γ-best fits.

Under the simplicity relation, setting $\gamma > 1 - \varepsilon$ again implies that the agent will discard the data generating process as a possible theory and subsequently hop between imposters. The implications of this switching between strategies are now not obvious. The agent chooses the simplest theories among those that provide γ-best fit. While the correct theory is not among them, it is not clear how well their predictions are correlated with the true data generating process. The following assumption adapts Assumption 4.3 to a tolerance for inaccuracy, and it allows us to compute the limit payoff for such values of γ.

Assumption 5 *For simplicity order* $\succsim^{LS,\gamma}$ *with* $\gamma > 1 - \varepsilon$ *and sufficiently large* n, $\left\{t \in B_{\succsim_h^{LS,\gamma}} \,|\, t(h) = 1 - \varepsilon\right\}$ *and* $\left\{t \in B_{\succsim_h^{LS,\gamma}} \,|\, t(h_n) = \varepsilon\right\}$ *are nonempty.*

Thus, we assume that the simplest theories are rich enough to contain theories that predict 0 and theories that predict 1. It is not obvious that the simplicity relation should have this property. If, for example, we observe the pattern 00000, it is not obvious that one of the simplest theories next will predict 1. However, when n is large, the actual data generating process has surely been discarded by order $\succsim_h^{LS,\gamma}$ and any theory amassing a likelihood above γ is surely a fluke. As a result, it is not obvious what *a priori* information, if any, should be brought to bear. One might then view this assumption as describing a potentially plausible worst-case scenario for the agent. The (omitted) proof of the following is then immediate:

Proposition 8 *Let Assumptions 1, 4.1–4.2 and 5 hold. Then* $P(d, \succsim^{LS,\gamma}) = \frac{1}{2}$.

Less drastic requirements than Assumption 5 will give similar but weaker results, with the key point being that setting $\gamma > 1 - \epsilon$ forces the agent to abandon any theory that sufficiently often predicts as does the true theory, in the process placing constraints on the payoff of which the agent can be assured.

4.3 Stability in Learning

One virtue of a simplicity order in a deterministic environment is that it prevents the agent from abandoning perfectly good theories. Setting $\gamma < 1-\varepsilon$ ensures that the agent will eventually retain the data generating process among the γ-best fits. This alone, however, does not ensure effective learning. Selecting the simplest γ-best fit leaves open the possibility that the agent may switch back and forth between theories that are simpler than the true one, where, at each period, one of the theories provides a γ-best fit, but fails to predict correctly. This is possible if the simple theories tend to be wrong precisely when they are used for prediction, but "catch up" in terms of the likelihood during periods in which they are not used for prediction. To see that this learner's nightmare might come true, consider the following.

Example 4 Fix $\gamma < (1 - \varepsilon)$ and let d be the data generating process. To simplify the presentation, but without losing any generality, assume that d predicts 1 in each period (with probability $1 - \varepsilon$).

We construct k theories, denoted by $t_1, \ldots, t_k$, which will be assumed the simplest: $t_1 \succ^S t_2 \succ^S \ldots \succ^S t_k$ and $t_k \succ^S t'$ for all $t' \notin \{t_1, ..., t_k\}$.

For concreteness, we describe the theories by an algorithm. For $n = 0$, $t_i(h_0) = 1$ for all $i \leq k$. For $n > 0$, given history h_n, every t_i ($i \leq k$) computes the predictions of all t_j ($j \leq k$, $j = i$ included) for all sub-histories h_m of h_n (for all $m < n$). By induction, this is a computable task. Next, each t_i computes $l(t_j, h_n)$ for all $j \leq k$. If none of them has a likelihood $l(t_j, h_n) \geq \gamma$, t_i predicts 1. Otherwise, t_i finds the simplest of the theories in $\{t_1, ..., t_k\}$ with $l(t_j, h_n) \geq \gamma$. If it is itself, it predicts 0; otherwise, it predicts 1.

Observe that each theory $\{t_1, ..., t_k\}$ basically performs the same algorithm, which simulates the calculations of all previous periods, and halts by induction. The difference between the predictions of the different theories in $\{t_1, ..., t_k\}$ arises only out of the very last step of the algorithm, in case some of them obtain a likelihood value above the threshold.

Observe also that in each period, at least $k-1$ of the theories $(t_1, \ldots, t_k)$ will produce a prediction matching that of d, and—if and only if some reach the appropriate likelihood threshold—one of these theories will dissent. Let ε_n be the proportion of realized 0's up to time n. The collective number of correct predictions among the k theories $(t_1, \ldots, t_k)$ in history h_n will thus be at least

$$[(1-\varepsilon_n)(k-1)]n,$$

where ε_n gets arbitrarily close to ε with arbitrarily large probability as n gets large. Hence, a lower bound on the number of correct predictions, among the k theories $(t_1, \ldots, t_k)$ over periods $0, \ldots, n-1$ is given by

$$[(1-\varepsilon-\delta)(k-1)]n$$

for some $\delta > 0$. We can choose n^* sufficiently large that

$$\delta < \frac{(1-\varepsilon)-\gamma}{2}$$

and then k sufficiently large that, for all $n > n^*$,

$$\left[(1-\varepsilon-\frac{(1-\varepsilon)-\gamma}{2})(k-1)\right]n > k\gamma n, \tag{10}$$

or

$$\frac{k-1}{k}\left(\frac{1-\varepsilon+\gamma}{2}\right) > \gamma.$$

(Since $1-\varepsilon > \gamma$, such a k exists.) From (10), we see that the theories $(t_1, \ldots, t_k)$ must have collectively amassed at least $k\gamma n$ correct predictions for any $n > n^*$, ensuring that at least one of them must have at least γn correct predictions, and hence a likelihood of at least $\theta(\gamma)$. As a result, one of these theories will be used for prediction in every period $n > n^*$, and by definition predicts that outcome which appears with probability ε under the data generating process d. Hence, the agent's payoff converges to ε. ■

It may appear as if the theories $(t_1, \ldots, t_k)$ in Example 4 are hopelessly special, tied closely to the structure of the true data generating process, and hence that the example is simply a curiosity. While we make no claims for the realism of the example, it is important to note that the simplicity order may include as relatively simple a multitude of collections $(t_1, \ldots, t_k)$,

corresponding to different data generating processes, with the likelihoods of those that do not match the actual data generating process falling until they are irrelevant, and then the calculations of the example becoming relevant. While still delicate, the phenomenon in the example is thus not as special as it may first appear. If we are to achieve a general result, we must have some additional structure.

There are likely to be many ways of addressing this problem. Our intuition provides one alternative: suppose you have several experts to choose from. Surely, how well they explain past observations is a key criterion, and so is simplicity. But it would appear natural to rely on an expert who has been consistently successful at explaining the data, rather than on one who boasts a great likelihood only at the present moment.

Formally, let there be given $\gamma \leq 1-\varepsilon$ and $k \geq 1$. For a theory t and history $h \in H_n$, $n \geq k$, define

$$\Gamma_{\gamma,k}(t) = \sum_{j=k}^{n} \delta_j,$$

where

$$\delta_j = \begin{cases} 1 & if \quad l\left(t, h_j\right) \geq \theta(\gamma) \\ 0 & if \quad l(t, h_j) < \theta(\gamma) \end{cases}$$

(where h_j is the j-th prefix of h). Next, define the relations $\succsim_h^{LS,\gamma k}$ for $h \in H$ as follows.

$$t \succsim_h^{LS,\gamma k} t' \iff \begin{cases} & [\Gamma_{\gamma,k}(t) > \Gamma_{\gamma,k}(t')] \\ or & [\Gamma_{\gamma,k}(t) = \Gamma_{\gamma,k}(t') \quad and \quad t \succsim^S t']. \end{cases}$$

Thus, a maximizer of $\succsim_h^{LS,\gamma k}$ has to be a theory that has obtained an average log-likelihood of at least $\theta(\gamma)$ as often as possible over the past consecutive $(n-k+1)$ periods. If there are several theories that obtained this likelihood threshold for the entire period, the maximizer has to be (one of) the simplest among them. If no theory has done as well as $\theta(\gamma)$ for $(n-k+1)$ periods, $\succsim_h^{LS,\gamma k}$ selects the simplest among those that have achieved at least $\theta(\gamma)$ for at least $(n-k)$ periods out of the past $(n-k+1)$ periods, and so forth.

Clearly, the choice of the parameters γ and k allows a wide range of relations $\left(\succsim_h^{LS,\gamma k}\right)$. What should be the values of γ and k and how are they

determined?[18]

We discuss two possibilities for determining γ and k. First, we assume that Evolution "programs" the tolerance for inaccuracy (γ) and the preference for stability (captured by k) into the reasoner's preferences, and, second, we endogenize this optimization process to be the result of a dynamic selection by the agent himself.

4.4 Evolutionarily Determined Tolerance

How much inaccuracy should the reasoner be willing to tolerate? The critical value $1-\varepsilon$ builds sufficient tolerance for inaccuracy into the agent's choices as to ensure effective learning:

Proposition 9 *Under Assumption 4, for every simplicity relation $\succsim^S$ and for every $d \in D$, $P(d, \succsim^{LS,\gamma k}) \to (1-\varepsilon)$ as $\gamma \nearrow 1-\varepsilon$ and $k \to \infty$.*

We thus find that, in the presence of randomness, augmenting the preference for simplicity with a preference for stability enhances the agent's payoff. Indeed, we tend to trust experts who have *always* provided good explanations more than experts who have *sometimes* provided good explanations. Even if two experts, or theories, reach the same level of goodness of fit at present, a better history may well be a reason to prefer one over the other.

Observe that one cannot do away with the preferences for simplicity and rely on stability alone. In the absence of a preference for simplicity, for every history h_n there exists a theory t_n such that $l(t_n, h_j) = \theta(1)$ for *every* $j \leq n$. Such a theory would maximize the likelihood function for each prefix of the history h_n, and would therefore be chosen for prediction. Thus the preference for stability alone does not provide a safeguard against overfitting the data by choosing a theory post-hoc.

4.5 Reasoned Tolerance

Section 4.4 assumed that Evolution looks for an optimal tolerance level γ and stability parameter k to equip the reasoner with the "right" preferences that can asymptotically select the correct theory. We can alternatively endow the agent with preferences that perform the same job.

[18]Notice that it makes no sense to insist on stability if one sets $\gamma > 1-\varepsilon$, since we know that no theory can long sustain a likelihood above $1-\varepsilon$.

Proposition 10 *Let Assumption 4 hold. For every simplicity order $\succcurlyeq^S$ there exists a relation $\succsim^{S*}$, independent of ε, such that*

(i) for every $d \in D$, we have $P(d, \succsim^{S*}) = 1 - \varepsilon$

and

(ii) for every $t, t' \in T$, for large enough n, if $\Gamma_{\gamma,k}(t) = \Gamma_{\gamma,k}(t')$, then

$$t \succsim^{S*} t' \qquad \Longleftrightarrow \qquad t \succcurlyeq^S t'.$$

The agent who implements $\succsim^{S*}$ engages not only in learning but also in meta-learning. This agent selects theories that provide a γ-best fit and that are simple, but at the same time, he observes his own learning process and learns from this process itself. Specifically, the agent looks at the choices he would have made for various levels of γ and asks, "What can I learn from the fact that for some levels of γ my learning process would have continued indefinitely, whereas for others I would have settled on a specific theory?" The fact that certain levels of γ do not let the agent converge on a given theory is taken to be an indication that this level is too high.

The parameter γ may be viewed as the agent's aspiration level for the degree of accuracy of the theory (in the sense of Simon [15]). We can imagine the agent setting a large value of γ in the hope of finding a theory that is quite close to the maximal likelihood one. However, if he finds that the search for such a theory does not result in a stable choice, and that he keeps bouncing around among theories no matter how large n is, then the agent may reduce his aspiration level γ. When γ is low enough, the agent will find a theory that has a higher degree of inaccuracy, but that can be chosen over and over again. This search for the optimal γ can be viewed as the search for the optimal aspiration level.

Remark 1 *The arguments behind Propositions 9 and 10 make it clear that nothing depends on the fixed error rate ε. Let D_* be the set of data generating processes with the property that, for every outcome h, there exists a pair $(\underline{\rho}, \overline{\rho} \in [0, 1/2) \times (1/2, 1]$, such that*

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T_+(h(n))} \sum_{n=1}^{T-1} d_+(h_n) &= \overline{\rho} \\ \lim_{T\to\infty} \frac{1}{T_-(h(n))} \sum_{n=1}^{T-1} d_-(h_n) &= \underline{\rho}, \end{aligned}$$

where $d_+(h_n)$ equals $d(h_n)$ if the latter exceeds $1/2$ and is zero otherwise, $d_-(h_n)$ is analogous for values of $d(h_n)$ less than $1/2$, $T_+(h(n))$ is the number of times theory d has produced a prediction exceeding $1/2$ on the history h_n, and $T_-((n))$ is analogous for predictions less than $1/2$. We are thus assuming that the average error rate in the data generating process, when predicting either 1 or 0, converges (though not necessarily to the same limits). If this is not the case, there is no hope for the agent to identify the appropriate error rates for effective learning. Then arguments analogous to those giving Proposition 10 allow us to establish that for every simplicity order $\succcurlyeq^S$, there exists a strategy $\succsim^{S*}$ such that the agent's limiting payoff in periods in which a 1 is predicted approaches $\overline{\rho}$ and the agent's limiting payoff in periods in in which a 0 is predicted approaches $\underline{\rho}$.

5 Discussion

5.1 Smooth Trade-Offs

Our central result is that effective learning couples concerns about a theory's likelihood with an auxiliary criterion we have interpreted as simplicity. Studies of model selection in statistics and in machine learning often similarly suggest a trade-off between likelihood and simplicity that, unlike our lexicographic criterion, is reflected in a smooth objective function. For example, the Akaike Information Criterion (Akaike [1]) is given by

$$\log(L(t)) - 2k,$$

where $L(t)$ is the likelihood function of theory t and k is the number of parameters used in model t. Related to Kolmogorov's complexity measure (Kolmogorov [9, 10], Solomonoff [16], Chaitin [3]), the Minimal Message Length criterion (Wallace and Boulton [20], Rissanen [12]) suggests

$$\log(L(t)) - MDL(t),$$

where the $MDL(t)$ is the minimum description length of the theory t. (See also Wallace and Dowe [21] and Wallace [19].)

The general form of these measures is

$$\log L(t) - \alpha C(t), \tag{11}$$

where $C(t)$ is a complexity function (cf. (5)) and α a constant determining the relative weights placed on the likelihood and on the complexity of the theory. Gilboa and Schmeidler [5] offer an axiomatization of this criterion. In their model the reasoner has an order over theories given data, akin to $\succsim_h$ in our case. Certain axioms on the way theories are ranked by this relation for different histories h imply an additive trade-off between the log-likelihood and a parameter of the theory that may be interpreted as its measure of complexity.

We cannot apply (11), designed to evaluate theories given a fixed set of data, directly to our setting. As we have noted, the likelihood $L(t)$ inevitably declines to zero and hence its log decreases without bound as observations accumulate. This ensures that complexity considerations will eventually play no role in the analysis. We accordingly examine

$$l(t,h) - \alpha C(t), \tag{12}$$

ensuring that likelihood and complexity considerations remain on a common footing.[19]

We can draw a connection between smooth measures such as (12) and our lexicographic criterion. Fix a complexity function $C(t)$ and parameter α, and let $\succsim^\alpha$ be the resulting order over theories induced by (12). How does $\succsim^\alpha$ compare to $\succsim^{LS}$?

To simplify the discussion, let us restrict attention to a set of data generating processes $D_\varepsilon^C \subset D_\varepsilon$ with the property that for any $d, d' \in D_\varepsilon^C$, the average log likelihood ratio $l(d', h_n)$ converges with probability one, when the data generating process is d. If we did not do this, $\succsim^\alpha$ could fall prey to instability of the type presented in Example 4, and would have to be supplemented by the type of stability criterion presented in Section 4.3 to be effective. Doing so would be straightforward, but would clutter the argument.

Proposition 11 *Let $D = T \subset D_\varepsilon^C$ be countable. Then*

$$\lim_{\alpha \to 0} P(d, \succsim^\alpha) = 1 - \varepsilon.$$

For a fixed α, the criterion $L(t) - \alpha C(t)$ restricts attention to a finite subset of D_ε^C as possible maximizers of $L(t) - \alpha C(t)$, since a theory that

[19] In so doing, we move close to criteria such as the Schwarz Information Criterion (also known as the Bayesian Information Criterion Schwarz [14]), which retains the additive trade-off but uses a complexity measure that depends on the number of observations.

is too complex can never amass a likelihood value large enough to exceed the value $L(t) - \alpha C(t)$ attained by the simplest theory. Among this finite set, no theory can consistently achieve a likelihood above $1 - \varepsilon$. If α is too large, this finite set will exclude the data generating process itself, and all of the eligible theories may well fall short of likelihood $1 - \varepsilon$. Smaller values of α will not exclude the data generating process a priori, but may still lead to the selection of a simpler theory and an attendant likelihood loss. As α gets arbitrarily small, we can be assured that the data generating process is encompassed in the set of eligible theories and that very little likelihood is sacrificed in the interests of simplicity, leading to a payoff approaching $1 - \varepsilon$.

Notice, however, that $P(d, \succsim^0) = P(d, \succsim^L)$, and hence $P(d, \succsim^0)$ equals $1/2$ (given Assumptions 1 and 4.3). In addition, we cannot say a priori how small α must be in order to ensure that $P(d, \succsim^\alpha)$ is close to $1 - \varepsilon$. We thus need to make α arbitrarily close to zero, without actually equalling zero. This is just what our lexicographic criterion does. We can accordingly view the lexicographic criterion as the limiting case of the smooth criteria that have been offered in the literature.

5.2 Bayesianism

We have observed that the Bayesian approach bears some similarity to "simplicism", i.e., to the simplicity-preferring approach. How do we compare the two?

The Bayesian approach requires computational capabilities that are much more demanding than the simplicistic one. The Bayesian approach requires a quantification of beliefs, and a comparison between the likelihood of every two sets of theories, not only of particular ones. This is particularly demanding in case there are infinitely many theories. Another complication arises out of the fact that a theory typically has many possible representations. For example, consider the theory t, "predict always 0" and the addition of the theory t', "At odd periods predict 0, and at even ones predict 0". A Bayesian reasoner would need to observe that the two are equivalent, and to assign both of them the same probability that he would have assigned t if it were the only one listed. That is, when contemplating a series of possible theories, say, described as Turing machines, a Bayesian reasoner is implicitly assumed to know which are observationally equivalent. In the case of computable theories, the task of verifying their equivalence is not computable itself.

By contrast, a simplicistic agent has no such concerns. Since he uses only

one theory for prediction, he need not conceive of all possible theories, let alone to verify their independence. Should the simplest theory happen to be equivalent to another, less simple theory, the predictions generated by a simplicity-liking reasoner would not change. Similarly, if a theory is "split" into several theories (say, seemingly more specific ones), all theories might still be considered in the set of candidate theories, and the reasoner need not subject this set to any pre-processing to guarantee their non-equivalence.

5.3 Stability and Simplicity

We found that, in the presence of uncertainty, stability needs to be added to likelihood and simplicity as a theory selection criterion. We could reduce the preference for stability to a preference for simplicity if we adopt a broader conceptualization of a "theory". Suppose the agent must not only choose a theory at each period, but also an explanation for the way the period-by-period choice is made. An agent who sticks to a particular theory t from some period onwards will have a simpler pattern of choice than another who continually switches among theories. Hence, if we ask the agent to select a simple *meta*-theory, which explains why specific theories are selected at different periods, a stable selection will have a lower complexity than an unstable one.

5.4 Simplicity and Language

The measurement of simplicity depends on language. This fundamental insight was made all too clear by Goodman's [6] famous "grue-bleen" paradox. In this paper we did not delve into this issue because our results do not depend on the particular measure of complexity one uses. Specifically, any enumeration of theories is sufficient to define a collection (due to equivalences) of simplicity relations. Moreover, taking a computational point of view, the differences between the simplicity rankings of different theories can be bounded (cf. Solomonoff [16]).[20]

[20]Consider two languages, L_1 and L_2. If the languages are computationally equivalent, there are compilers that can translate a program from L_1 to L_2 and vice versa. It is certainly possible that in L_1 the minimal description length (MDL) of a theory t, $MDL_1(t)$, is shorter than that of t', $MDL_1(t')$, while the converse is true in language L_2. But the MDL of each theory in language L_1 cannot exceed its MDL in language L_2 plus the length of the compiler translating from L_1 to L_2. The converse is also true, and this means that

5.5 Statistics and Evolutionary Psychology

Statistics and evolutionary psychology are not typically seen as closely related disciplines. Researchers in statistics and in machine learning tend to seek optimal inference and learning techniques. Evolutionary psychologists seek to explain psychological phenomena as solutions to realistic constrained optimization problems faced by an organism. Both fields seek optimal solutions, but they assume different constraints. Nonetheless, similar types of reasoning have been independently developed in both domains. The preference for simplicity can be viewed as a case in which the natural preference of people in reasoning coincides with the normative considerations of statisticians.

Similar concurrences can be found in other examples. Techniques of nearest-neighbor classifications are similar to categorization by examples. The same is true of kernel-based probabilities and "exemplar learning." Moreover, Gayer [4] points out that the bias of kernel methods familiar in the statistical literature can also serve as an explanation of the "distorion" of probabilities in Prospect Theory (Kahneman and Tversky [8]). Thus, both the successes and the failures of certain statistical techniques can also be found in the human mind. We find it encouraging that evolution has endowed the human mind with some of the patterns of reasoning developed in statistics and in machine learning.

6 Appendix: Proofs

Proof of Proposition 1.1. Fix $d \in D = T$. There are finitely many theories in $S(d) \equiv \left\{t \in T \,|\, t \succsim^S d\right\}$, i.e., that are as simple as or simpler than d. Choose some $t \in S(d)$ and suppose that t and d are not observationally equivalent, meaning that they do not generate identical outcomes $((x_0, y_0), ..., (x_n, y_n), ...)$. Then at some period n theory t will be refuted, i.e., the data generating process will produce a history $h_n = ((x_0, y_0), ..., (x_{n-1}, y_{n-1}), x_n)$

---

there is a bound on the difference between the MDL of a theory in the two languages. That is, there exists c such that

$$|MDL_1(t) - MDL_2(t)| < c$$

for all t. Hence, if theory t has a sufficiently shorter MDL in L_1 than does theory t', this ranking will have to be preserved in language L_2.

for which $L(t, h_n) = 0$ and hence $d \succ_h^{LS} t$. Applying this argument to the finitely many theories in $S(d)$, there must exist a finite time n' by which either theory d is chosen by $\succsim_{h_{n'}}^{LS}$ or some element $t \in S(d)$ is chosen by $\succsim_{h_{n'}}^{LS}$ that is observationally equivalent to d. Thereafter, $p(d, t_n, h_n) = 1$ holds. This yields $P(d, \succsim^{LS}) = 1$ ■

Proof of Proposition 1.2. Assumption 2.3 ensures that, for every history h_n there are theories $t \in D$ consistent with h_n, that is, $L(t, h_n) = 1$. Consider

$$B_{\succsim_h^L} = \{\, d \,|\, L(d, h) = 1 \,\}.$$

For any finite continuation of h_n there is a theory $t \in B_{\succsim_h^L}$ that is consistent with this continuation. In particular, this is true for the history h_{n+1} generated from h_n and $y_n = 0$ (coupled with x_{n+1}) as well as for the history h'_{n+1} generated from h_n and $y_n = 1$ (coupled with x_{n+1}). Assumption 1 then ensures that the order $\succsim^L$ is equally likely to select a theory predicting $y_n = 0$ as it is to select a theory $y_n = 1$. Thus, the probability of making a correct prediction is 1/2, and hence $p(d, h_n, t_n) = 0.5$, regardless of the true process d. This establishes

$$P(d, \succsim_h^L) = 0.5.$$

We conclude that, for every simplicity relation $\succsim^S$, and for every data generating process d, the limit payoff under $\succsim^{LS}$ is 1, while it is only 0.5 under $\succsim^L$. Hence, $\succsim^{LS}$ strictly dominates $\succsim^L$. ■

Proof of Proposition 2. Consider an agent characterized by $\succsim^{LI}$ and suppose Fate has chosen theory d. If $P(d, \succsim^{LI}) < 1$, it must be that infinitely often, $p(d, t_j, h_j) = 0$. Hence, the agent infinitely often chooses a new theory but never chooses d. By Assumption 3, the probability that the agent chooses a new theory n times without choosing d is at most $(1 - \lambda(d))^n$. Since $\lim_{n\to\infty}(1 - \lambda(d))^n = 0$, the probability that $P(d, \succsim^{LI}) < 1$ is zero, and hence the expected value of $P(d, \succsim^{LI})$ is unity. ■

Proof of Proposition 3. The relation $T \times T$ guarantees a random choice (by Assumption 2.3), and hence this relation ensures an expected payoff of 0.5 at each period in which it is played. Thus, if $\succsim^{LS,\varepsilon} = T \times T$ for a long enough period, the average payoff converges to 0.5 with probability 1. Moreover, it does so at a rate proportional to $n^{-1/2}$. It follows that, with

probability 1, the sustained application of relation $T \times T$ leads to a period n at which the average payoff surpasses the threshold $0.5 - \varepsilon/\log n$, at which point $\succsim^{LS,\varepsilon}$ switches to $\succsim^{LS}$.

Suppose that Fate chooses $d \in T$. Since $\succsim^{LS,\varepsilon}=\succsim^{LS}$ infinitely often, $\succsim^{LS,\varepsilon}$ will eventually select d or a theory equivalent to d. Predictions will subsequently be perfect, ensuring that $\succsim^{LS,\varepsilon}$ will not revert to $T \times T$ and that $P(d, \succsim^{LS,\varepsilon}) = 1$.

If Fate chooses $d \notin T$, the average payoff at history h_n can drop no lower than $0.5 - \varepsilon/\log n - 1/n$ (obtained by coupling a history of length $n-1$ in which the payoff is precisely $0.5 - \varepsilon/\log(n-1)$ with one more incorrect observation) before $\succsim^{LS,\varepsilon}= T \times T$. Hence $P(d, \succsim^{LS,\varepsilon}) \geq 0.5$. Combining the two, we thus find that $P(d, \succsim^{LS,\varepsilon}) \geq P(d, \succsim^{L})$ for all $d \in D$, with strict equality for every $d \in T$. ■

Proof of Proposition 4. Let $\succsim$ be computable. Then there is a Turing machine τ that implements $\succsim$ by, for any history h, computing a maximizer of $\succsim$ from the set D^H. Let d simulate the machine τ, for any history h, finding the maximizer t_h that the agent will use for prediction, and then generating prediction 1 if $t_h(h) \leq 0.5$ and 0 if $t_h(h) > 0.5$. A deterministic t will result in a payoff of 0. The maximal payoff for the agent at each period is 0.5, obtained by the random prediction $t_h(h) = 0.5$. ■

Proof of Proposition 5. The basic idea is to construct the relation $\succsim$ by combining the underlying simplicity order $\succcurlyeq$ with the time complexity of the machine.

Let $D^T = \{t_1, t_2, ...\}$ be the class of all Turing machines, including those that always halt and those that do not halt for certain inputs $h \in H$. There is no difficulty in writing a machine that generates D^T, or, equivalently, that can accept $i \geq 1$ as an input and, after a finite number of steps, provide the description of t_i.

Assume we are given a history h and we wish to select a theory that has high likelihood and that halts for h. When considering a machine t, we thus need to determine whether it fits the data, namely whether $L(t, h_n) = 1$ (taking $L(t, h_n) = 0$ if the machine fails to halt for any prefix of h_n), and we need to compute its prediction for y_n, or $t(h_n)$, taking into account the possibility that it may not halt when making this prediction. That is, we need to know the result of $n+1$ computations of t_i, each of which may not

halt.

Let $C : D \to \mathbb{N}$ be a computable complexity function for the underlying simplicity order $\succcurlyeq^S$, so that

$$C(t) \leq C(t') \qquad \Longleftrightarrow \qquad t \succcurlyeq^S t'.$$

Define $c : D \times H \to \mathbb{N} \cup \{\infty\}$ to be the length of computation, that is, $c(t,h) \in \{1,2,...,\infty\}$ is the number of steps that t takes to compute where h is its input. Next define a function $C^* : D \times H \to \mathbb{R}_+ \cup \{\infty\}$ by

$$C^*(t,h) = C(t) + \frac{1}{n^2}\sum_{j=0}^{n} c(t,h_j)$$

where $t \in D$, $h \in H_n$ and h_j is the j-th prefix of h. Using this function, we define our candidate relation over theories:

$$t' \succsim_h t \qquad \Longleftrightarrow \qquad \begin{cases} & L(t',h) > L(t,h) \\ or & [L(t',h) = L(t,h) \;\; and \;\; C^*(t',h) \leq C^*(t,h)] \end{cases}.$$

We argue that it is a computable task to find a maximizer of $\succsim_h$ from among those machines that halt on history h, and that this maximizer will have likelihood one. First observe that for every h there exists a machine t such that $L(t,h_n) = 1$ and $C^*(t,h_n) < \infty$. To see this, it suffices to consider a machine t that generates history h_n irrespective of the data. For any history longer than n, the machine can generate 0. This takes a computation time $c(t,h) = O(n)$. By construction, $t \in D_0^B$. Since this machine appears somewhere in the enumeration corresponding to $\succcurlyeq^S$, we have $C(t) < \infty$ and hence $C^*(t,h) < \infty$.

Given $C^*(t,h)$, there are finitely many machines t' with $C(t') \leq C^*(t,h)$, and therefore only finitely many machines that can beat t according to $\succsim$. Each of these has to be simulated only a bounded number of steps, $C^*(t,h)$, to see if, indeed, it gives $L(t',h_n) = 1$ and a lower value for $C^*(t',h)$.

Note that for all $d \in D_0^B$, $c(t,h_n) \leq K(d)$ and

$$\frac{1}{n^2}\sum_{j=0}^{n} c(t,h_j) \leq \frac{1}{n^2} nK(d) \to 0$$

hence,

$$C^*(t,h) \to C(t).$$

Now consider $d, d' \in D_0^B$ with $d \succ^S d'$ and hence $C(d) \leq C(d')$. Then for all sufficiently large n, $C^*(d, h_n) < C^*(d', h_n)$, and hence $L(d, h_n) = L(d', h_n) \Rightarrow d \succ_h d'$. This establishes (5.2).

We now turn to (5.1), namely that $P(\succsim, d) = 1$ for every $d \in D_0^B$. For $t' \succsim_h d$ to hold, we must have $L(t', h) = 1$ and $C(t') \leq C(d)$. An argument analogous to that of the proof of Proposition 1 ensures that at some point, d or a theory equivalent to it is found, and from that point on only such theories (predicting $d(h)$ for every h) can be maximizers of $\succsim$. Hence the agent makes perfect predictions and obtains $P(\succsim, d) = 1$. ■

Proof of Proposition 9. Fix a data generating process d. Assume that γ satisfies $\theta(\gamma) = \theta(1-\varepsilon) - \delta$ for $\delta > 0$. For any $\eta > 0$, there exists k such that, with probability $1 - \eta$ at least, for all $n \geq k$,

$$l(d, h_n) > [(1-\varepsilon)\log(1-\varepsilon) + \varepsilon \log \varepsilon] - \delta = \theta(1-\varepsilon) - \delta = \theta(\gamma).$$

Thus, from period k on, it is likely that the correct theory d is among the γ-maximizers of $l(\cdot, h_n)$. If d is the maximizer of $\succsim^{LS,\gamma k}$ used for prediction, a payoff of $(1-\varepsilon)$ is guaranteed. We wish to show that, if another theory is used for prediction, it cannot be much worse than d itself.

Let us condition on the probability $1 - \eta$ event that for every $n > k$, $l(d, h_n) > \theta(\gamma)$. If a theory $t \neq d$ is used for prediction at period $n \geq k$, then it must be the case that (i) t is a γ-best fit for all periods $j = k, ..., n$; and (ii) $t \succsim^S d$. Hence, for each period $n > k$, there are only a finite number of theories satisfying conditions (i) and (ii), of which the simplest will be chosen. Moreover, the set of such theories is decreasing in n (since a theory whose likelihood ratio drops below γ is subsequently disqualified). Eventually, a period n' will be reached such that some theory t (possibly d) satisfying (i) and (ii) will be used in all subsequent periods. Let $n > n'$, and let α be the proportion of times, up to n, that t made the correct prediction. Then, since

t is a γ-best fit at n, we have

$$\begin{aligned} l(t,h) &= \alpha \log(1-\varepsilon) + (1-\alpha)\log\varepsilon \\ &= \alpha\left[\log(1-\varepsilon) - \log\varepsilon\right] + \log\varepsilon \\ &= \alpha \log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon \\ &\geq \theta(\gamma) \\ &= \theta(1-\varepsilon) - \delta \\ &= (1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log\varepsilon - \delta \\ &= (1-\varepsilon)\left[\log(1-\varepsilon) - \log\varepsilon\right] + \log\varepsilon - \delta \\ &= (1-\varepsilon)\log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon - \delta. \end{aligned}$$

This gives

$$\alpha \log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon \geq (1-\varepsilon)\log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon - \delta$$

or

$$[\alpha - (1-\varepsilon)]\log\frac{1-\varepsilon}{\varepsilon} \geq -\delta$$

that is,

$$\alpha \geq (1-\varepsilon) - \frac{\delta}{\log\frac{1-\varepsilon}{\varepsilon}}.$$

Intuitively, the payoff obtained by predicting according to t cannot be much lower than $(1-\varepsilon)$. Taking into account the probability of convergence by time k we get

$$P(d, \succsim^{LS,\gamma k}) \geq (1-\eta)\left[(1-\varepsilon) - \frac{\delta}{\log\frac{1-\varepsilon}{\varepsilon}}\right],$$

which converges to $(1-\eta)(1-\varepsilon)$ as $\delta \searrow 0$. Finally, increasing k results in decreasing η to any desired degree, and the result follows. ■

Proof of Proposition 10. The basic idea is have the agent simulate the choices of theories that would have corresponded to $\succsim^{LS,\gamma k}$ for different values of γ and of k. For values of γ larger than $1-\varepsilon$, the agent will find that the maximizers of $\succsim^{LS,\gamma k}$ keep changing, indicating that γ is too high. For values of γ that are lower than $1-\varepsilon$, the agent will find theories that get selected

asymptotically, an indication that γ might be too low. By refining the search for γ, while simultaneously gathering more observations, the reasoner will approach $1-\varepsilon$ and make predictions according to the correct theory.

We make these ideas precise in the form of a reasoning algorithm that is simple, but makes no claims to efficiency. At stage n the reasoner considers as possibilities for γ all values in

$$\Gamma_n = \left\{ \frac{r}{2^n} \,|\, r = 0, 1, ..., 2^n \right\}.$$

Given n, define $k = \llcorner n/2 \lrcorner$. For each $\gamma \in \Gamma_n$, and for each $m = k, ..., n$, the reasoner finds all the maximizers of $\succsim_{h_m}^{LS,\gamma k}$ (to make this an algorithm, we need to assume that an oracle can perform this task). Denote the set of maximizers for each γ by $M(m, k, \gamma)$. This is a finite set, due to the agent's preference for simplicity. Then, for each γ, define

$$M^*(n, \gamma) = \cap_{k \leq m \leq n} M(m, k, \gamma).$$

Thus, $M^*(n, \gamma)$ contains precisely those theories that have been among the "γ-best" theories for the past $n/2$ periods.

If $M^*(n, \gamma) = \varnothing$ for all $\gamma \in \Gamma_n$, define $\succsim_{h_n}^{S*} = D \times D$. In this case all theories are equivalent in terms of $\succsim_{h_n}^{S*}$, and the reasoner's choice will be arbitrary.

If, however, $M^*(n, \gamma) \neq \varnothing$ for some $\gamma \in \Gamma_n$, let γ_n be the maximal such value in Γ_n, and define

$$t \succsim_{h_n}^{S*} t' \quad \Longleftrightarrow \quad \left\{ \begin{array}{rrcll} & t & \in & M^*(n, \gamma_n) & and \quad t' \notin M^*(n, \gamma_n) \\ or & t, t' & \in & M^*(n, \gamma_n) & and \quad t \succsim^S t' \\ or & t, t' & \notin & M^*(n, \gamma_n). & \end{array} \right.$$

That is, the $\succsim^S$-simplest theories in $M^*(n, \gamma_n)$ are considered to be the "best" theories and one of them will be used for prediction.

To see that the definition of $\succsim^{S*}$ satisfies the desired properties, observe that, by the proof of Proposition 9, if $\gamma > 1-\varepsilon$, $M^*(n, \gamma) = \varnothing$ for large n. For $\gamma < 1-\varepsilon$, $d \in M^*(n, \gamma)$ for large n. As $n \to \infty$, the minimal γ for which $M^*(n, \gamma) \neq \varnothing$ converges to $1-\varepsilon$, and d is among the maximizers of $\succsim^{S*}$. We then repeat the argument of Proposition 9, by which any theory $t \neq d$ such that $t \in M^*(n, \gamma)$ obtains a payoff that converges to $(1-\varepsilon)$ as $\gamma \nearrow 1-\varepsilon$. ■

Proof of Proposition 11. Fix a complexity function $C(t)$, a value $\alpha > 0$, and a data generating process d^*. Let $\hat{d} \in \arg\min_{d \in D_\varepsilon^C} C(d)$. Then no theory d for which $\theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})$ will ever be chosen by the relation $\succsim^\alpha$, no matter what the history. The agent's choice of theory in each period will thus be drawn from the finite set $D_\varepsilon^C(\alpha) \equiv \{d \in D_\varepsilon^C : \theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})\}$.

For sufficiently small α, the data generating process d^* is contained in $D_\varepsilon^C(\alpha)$. In addition, with probability 1, the limit $\lim_{n\to\infty} l(d, h_n)$ exists for all $d \in D_\varepsilon^C(\alpha)$. Since this set is finite, with probability 1, the agent's choice of theory becomes constant across periods, being the maximizer over $D_\varepsilon^C(\alpha)$ of

$$\lim_{n\to\infty} l(d, h_n) - \alpha C(d).$$

But since $d^* \in D_\varepsilon^C(\alpha)$ for small α, the agent's payoff is bounded below by

$$\lim_{n\to\infty} l(d, h_n) - \alpha C(d) = \theta(1 - \varepsilon) - \alpha C(d^*).$$

Taking α to zero then gives the result. ■

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Nabil Ibraheem Al-Najjar. Decision makers as statisticians. Technical report, Northwestern University, 2008.

[3] Gregory J. Chaitin. On the length of programs for computing binary sequences. *Journal of the Association for Computing Machinery*, 13(4):547–569, 1966.

[4] Gabrielle Gayer. Perception of probabilities in situations of risk: A case-based approach. Mimeo, 2006.

[5] Itzhak Gilboa and David Schmeidler. Likelihood and simplicity: An axiomatic approach. Mimeo, Tel Aviv University, 2008.

[6] Nelson Goodman. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, Massachusetts, 1954.

[7] John E. Hopcraft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison Wesley, Reading, Mass., 1979.

[8] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[9] Andrei. N. Kolmogorov. Three approaches to the quantitative definition of information. *Probability and Information Transmission*, 2(1):4–7, 1965.

[10] Andrei. N. Kolmogorov. On tables of random numbers. *Theoretical Computer Science*, 207(6):387–395, 1998 (originally 1963).

[11] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1996.

[12] Jorma Rissanen. Modelling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[13] Bertrand Russell. *History of Western Philosophy*. Routledge, London, 2004. Originally 1946.

[14] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[15] Herbert Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–118, 1955.

[16] Ray J. Solomonoff. A formal theory of inductive inference I,II. *Information Control*, 7(1,2):1–22, 224–254, 1994.

[17] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[18] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[19] Christopher S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, New York, 2005.

[20] Christopher S. Wallace and D. M. Boulton. An information measure of classification. *The Computer Journal*, 13:185–194, 1968.

[21] Christopher S. Wallace and David L. Dowe. Minimal message length and Kolmogorov complexity. *The Computer Journal*, 42:270–283, 1999.

[22] Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Kegan Paul, London, 1923.