Dynamics of Inductive Inference in a Unified Framework[1]

Itzhak Gilboa,[2] Larry Samuelson,[3] David Schmeidler[4]

September 12, 2010

Abstract

We present a model of inductive inference that includes, as special cases, Bayesian reasoning, case-based reasoning, and rule-based reasoning. This unified approach allows us to examine how the various modes of inductive inference can be combined and how their relative weights change endogenously. We establish conditions under which an agent who does not know the structure of the data generating process will decrease, over the course of her reasoning, the weight of credence put on Bayesian vs. non-Bayesian reasoning. We show that even random data can make certain theories seem plausible and hence increase the weight of rule-based vs. case-based reasoning, leading the agent in some cases to cycle between being rule-based and case-based.

[1]We thank Dirk Bergemann, Eddie Dekel, Drew Fudernberg, Gabi Gayer, Offer Lieberman, George Mailath, the editors and three referees for comments and suggestions. Financial support from the National Science Foundation (grants SES-0549946 and SES-0850263) is gratefully acknowledged.

[2]Tel-Aviv University, HEC, Paris, and Cowles Foundation, Yale University.

[3]Department of Economics, Yale University.

[4]The Ohio State University and Tel-Aviv University.

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

September 12, 2010

Contents

1 Introduction **1**

2 The Framework **3**

3 Special Cases **8**
3.1 Bayesian Reasoning 8
3.2 Case-Based Reasoning 10
3.3 Rule-Based Reasoning 13
3.4 Combined models 15

4 Dynamics of Reasoning Methods **16**
4.1 Bayesian vs. non-Bayesian Reasoning 16
4.1.1 Assumptions 16
4.1.2 Result 19
4.1.3 Generalizations 23
4.1.4 When will Bayesianism Prevail? 24
4.2 Case-Based vs. Rule-Based Reasoning 27

5 Discussion **30**
5.1 Methods for Generating Hypotheses 30
5.2 Probabilistic Hypotheses 31
5.3 Single-Hypothesis Predictions 31
5.4 Decision Theory 32

6 Appendix: Proof of Theorem 1 **33**

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

September 12, 2010

1 Introduction

Consider an agent who each year is called upon to predict the price of oil over the subsequent year. To keep this illustration simple, suppose the agent need only predict whether the average price will be higher or lower than the previous year's price. We can imagine the agent working for a hedge fund that is interested in whether it should bet for or against an increasing price.

The agent does not make her oil-price prediction in a vacuum. Each year, her research staff compiles a long list of data potentially relevant to the price of oil, as well as a wealth of data identifying past values of the relevant variables and past oil prices. For our example, however, let us assume that the data include just two variables, namely a measure of the change in the demand for oil and a measure of the change in the severity of conflict in the Middle East. Each is assumed to take two values, indicating whether there has been an increase or decrease in the relevant measure. Each year the agent receives the current changes in demand and in conflict, examines the data from previous years, and then predicts whether the price will increase or decrease. How should the agent reason?[1]

The mode of reasoning most widely used in economic modeling is *Bayesian*. The agent first formulates the set of possible states of the world, where a state identifies the strength of demand, the measure of conflict, and the price of oil, in each year over the course of her horizon. The agent then formulates a prior probability distribution over this state space, identifying the relative weights the agent places on the various states. This prior distribution will reflect models and theories of the market for oil that the agent finds helpful, her analysis of past data and past events in this market, and any other prior information she has at her command. Once this prior has been formulated, the agent's predictions are a relatively straightforward matter of applying Bayes's rule, as new observations allow her to rule out some states and condition her probability distribution on the surviving states in order to make

[1] In personal conversation, a hedge fund principal indicated that his fund used all three methods of reasoning introduced in this section in predicting the likelihood of mortgage defaults.

new predictions.

Another mode of reasoning is *case-based*. Each year, the agent looks for similar cases in the past, and regards an outcome as more likely if it has materialized more often in more similar past cases. For example, she may argue that the current state of conflict in the Middle East is reminiscent of the state of affairs in 1991 or in 2003, and hence predict that there will soon be a war and an increase in the price of oil.

Finally, *rule-based* reasoning calls for the agent to base her predictions on regularities that she believes characterize the market for oil. For example, the agent may adopt a rule that any increase in the level of demand leads to an increase in the price of oil. Based on this and her expectation that the Chinese economy will continue to grow, the agent might reasonably predict that the price is about to rise.

The boundaries between the three modes of reasoning are not always sharp. Case-based and rule-based agents can update the probabilities they attach to the validity of various analogies or rules in light of their experience, much as would a Bayesian. A Bayesian will base her choice of a prior probability on analogies to similar past cases, as well as on general rules that she has observed. To make one boundary precise, we say that reasoning is "Bayesian" if all past analogies, regularities, and other prior information can be summarized in a prior probability distribution over the possible remaining histories, with all subsequent reasoning captured by standard Bayesian updating.

This paper presents a framework that unifies these three modes of reasoning (and potentially others), allowing us to view them as special cases of a general learning process. The agent attaches weights to hypotheses. Each hypothesis is a set of states of the world, which captures a way of thinking about how events in the world will develop, and the associated weights captures the relative influence that the agent attaches to the various hypotheses. To generate a prediction, the agent sums the weight of all hypotheses that necessitate each possible outcome, and then ranks outcomes according to their associated total weights. Learning is performed simply by ruling out hypotheses that have been proven wrong. In the special case where each hypothesis consists of a single state of the world, our framework is the standard Bayesian model, and the learning algorithm is equivalent to Bayesian updating. Employing other hypotheses, which include more than a single state each, we can capture other modes of reasoning, as illustrated by simple examples of case-based and of rule-based reasoning.

The unified framework allows us to examine the relationship between the various modes of reasoning, identify their differences, and delineate the scope of their applicability. Moreover, using this framework, we can analyze the dynamics of the weights assigned to the different reasoning modes as data accumulate. We first compare Bayesian to non-Bayesian modes of reasoning and identify conditions under which Bayesian reasoning will give way to other modes of reasoning. This result describes the outcome of the reasoning process, without taking a stance regarding its effectiveness or optimality. One could use our framework to add assumptions about the data generating process, in order to derive further results about the optimality of reasoning modes relative to the world to which they are applied, but our focus in this paper is on the question of which mode of reasoning will emerge as dominant in the agent's mind. We then use the assumptions behind this result to distinguish between situations in which the Bayesian approach is likely to be robust and situations in which it is not.

Finally, we discuss case-based and rule-based reasoning, showing that even random data can occasionally give rise to belief in a specific theory and hence in rule-based reasoning in general, until the theory is refuted and agents resort to case-based reasoning, potentially leading them to cycle between case-based and rule-based reasoning.

2 The Framework

At each period $t \in \{0, 1, \ldots, T-1\}$ there is a *characteristic* $x_t \in X$ and an *outcome* $y_t \in Y$. The sets X and Y are assumed to be finite and non-empty, with Y containing at least two possible outcomes.[2]

In predicting the price of oil, the characteristic x_t might identify the type of political regime and the state of political unrest in various oil-producing countries, as well as describe the extent of armed conflict in the Middle East, indicate whether new nuclear power plants have come on line or existing ones been disabled by accidents, describe the economic conditions of the major

[2]No conceptual problems arise in extending the analysis to infinite sets X and Y or an infinite number of periods. The sums over hypotheses that appear below would then be replaced by integrals. However, this requires the definition of an algebra, each element of which is a set of sets of states of the world. Specifying this algebra gives rise to a collection of technical complications that veil the message of the paper without having substantive implications.

oil importers, summarize climate conditions, and so on. In our simplified example, Y has only two elements, $\{0,1\}$, and each $x = (x^1, x^2) \in X$ has two components, each also taking values in $\{0,1\}$, with a 1 in each case indicating an increase in the relevant variable and a 0 indicating a decrease (or no change).

We make no assumptions about independence or conditional independence of the variables across periods. Our preferred interpretation is that this lack of structure reflects the agent's lack of knowledge about the data generating process—we are most interested in cases in which the agent has no certain knowledge that she can bring to bear on the prediction problem. For example, we do not think of statistical inference, in which the agent knows she faces a sequence of independent random variables from a fixed distribution, as our prime application. This is in keeping with our example of an agent who must predict long-term movements in the price of oil, rather than daily fluctuations of the price around a long-term trend.[3]

A *state of the world* ω identifies the characteristic and outcome that appear in each period t, i.e., $\omega : \{0, 1, ..., T-1\} \to X \times Y$. We let $(\omega_X(t), \omega_Y(t))$ denote the element (x_t, y_t) of $X \times Y$ appearing in period t given state ω, and let

$$\Omega = (X \times Y)^T$$

denote the set of states of the world. In our example, a state identifies the sign of changes in the strength of demand, the level of conflict, and the price of oil in each of the next T periods.

A period-t history

$$h_t(\omega) = (\omega(0), \ldots, \omega(t-1), \omega_X(t))$$

identifies the characteristics (sign of changes in the levels of demand and of conflict) and outcomes (sign of changes in the price of oil) that have appeared in periods 0 through $t-1$, as well as the period-t characteristic, given state ω. We let H_t denote all possible histories at period t, i.e., $H_t = \{h_t(\omega) \mid \omega \in \Omega\}$.

In each period t the agent observes a history h_t and makes a prediction about the period-t outcome, $\omega_Y(t) \in Y$. A *prediction* is a ranking of subsets in Y given h_t. In our example, $Y = \{0,1\}$, and the only interesting subsets to compare are those consisting of specific outcomes, $\{0\}$ and $\{1\}$. But in a

[3]When statistical learning is possible, we would be most interested in the unstructured learning process that remains after the agent has learned what she can from such inference.

richer model Y could consist of all possible prices of oil, and we would allow the agent to consider subsets of Y of the form "the price of oil will exceed \$100 per barrel" or "the price of oil will be below \$80 a barrel," and to rank some such subsets as being more likely than others. Hence, the agent may view a price of oil above \$100 as being more likely than a price under \$100, which is in turn more likely than a price of precisely \$110; or, she may view an increase in price as more likely than a decrease, and so forth.

Predictions are made with the help of hypotheses. A *hypothesis* is an event $A \subset \Omega$. It can represent a specific scenario, that is, a single state of the world, in which case $A = \{\omega\}$, and such hypotheses will suffice to capture Bayesian reasoning. However, hypotheses can contain more than one state, and thereby capture rules and analogies, as illustrated in the next section. In general, any reasoning aid one may employ in predicting y_t can be described by the set of states that are compatible with it. Let $\mathcal{A}$ denote the set of all hypotheses, that is $\mathcal{A} = 2^{\Omega}$.

The agent makes use of these hypotheses, and in particular assigns various weights of credence to them, with the help of a model. Formally, a *model* is a function $\phi : \mathcal{A} \to \mathbb{R}_+$, where $\phi(A)$ is interpreted as the weight attached to hypothesis A for the purpose of prediction. This model will bring to bear all of the prior information the agent has about the prediction problem.

For a subset of hypotheses $\mathcal{D} \subset \mathcal{A}$, we denote the total weight of credence commanded by these hypotheses by

$$\phi(\mathcal{D}) = \sum_{A \in \mathcal{D}} \phi(A).$$

For example, we might be interested in the total weight attached to all hypotheses that predict an increase in the price of oil.

To make predictions in period t, the agent first identifies, for any subset of outcomes $Y' \subset Y$, the set of hypotheses that have not been refuted by previous observations and that predict an outcome in Y'. She then adds the weights of credence attached to these hypotheses. The agent considers the set of outcomes Y' as more likely than the set Y'' if and only if the former attains a higher total weight of credence than the latter.

Formally, suppose the agent has observed history h_t in period t and considers the set of outcomes Y'. Then

$$[h_t] = \{\omega \in \Omega \mid (\omega(0), \ldots, \omega(t-1), \omega_X(t)) = h_t\}$$

is the set of all states that are compatible with the history h_t, that is, $[h_t]$ is the set of states whose period-t history matches h_t, with different states in this set corresponding to different possible future developments. Next,

$$[h_t, Y'] = \{\omega \in [h_t] \mid \omega_Y(t) \in Y'\}$$

is the set of all states that are compatible with the history h_t and with the next outcome being in the set Y'. A hypothesis A has not been refuted by history h_t if $A \cap [h_t] \neq \varnothing$. The set of hypotheses that have not been refuted by history h_t *and* predict an outcome in Y' is

$$\mathcal{A}(h_t, Y') = \{A \in \mathcal{A} \mid \varnothing \neq A \cap [h_t] \subset [h_t, Y']\}, \tag{1}$$

and hence the total weight assigned to Y' by all the unrefuted hypotheses at h_t is $\phi(\mathcal{A}(h_t, Y'))$.[4]

The agent's prediction is a ranking of the subsets of Y, with Y' considered more likely than Y'' if and only if

$$\phi(\mathcal{A}(h_t, Y')) > \phi(\mathcal{A}(h_t, Y'')). \tag{2}$$

It sacrifices no generality to assume that $\phi(\mathcal{A}) = 1$. Indeed, one may continually renormalize the function ϕ so that the total weight of all unrefuted hypotheses is 1 at each and every history h_t (unless, of course, this weight is zero), as is the case with Bayesian updating, without affecting the results—doing so would never reverse an inequality of the type given in (2). We do not follow this convention here to simplify the formulae and to avoid the need to deal with zero denominators separately.

Intuitively, one may think of each hypothesis A as an expert, who argues that the state of the world has to be in the event A. The weight $\phi(A)$ is a measure of the expert's reliability in the eyes of the agent. The agent listens to the forecasts of all experts and, when comparing two possible predictions Y' and Y'', chooses the prediction that commands higher total support from the experts. When an expert is proven wrong, he is asked to leave the room and his future forecasts are ignored.

The use of states of the world to represent possible outcomes is standard in decision theory, as is the summation of a function such as ϕ to capture beliefs,

[4]Observe that the hypotheses $\varnothing$ and Ω are never included in $\mathcal{A}(h_t, Y')$ for any $Y' \subsetneq Y$. The impossible hypothesis $\varnothing$ is not compatible with any history h_t, whereas the certain hypothesis Ω is tautological at every history h_t.

and the elimination of hypotheses that have been proven wrong. The only departure we have taken from the familiar framework of Bayesian updating is to allow hypotheses that consist of more than one state.[5] To confirm this, Section 3.1 shows that if we restrict attention to singe-state hypotheses, then we have precisely the familiar framework of Bayesian reasoning. Expanding the framework to encompass multi-state hypotheses is necessary if we are to capture case-based and rule-based reasoning (cf. Sections 3.2 and 3.3). Our framework thus captures case-based and rule-based reasoning by making the most obvious, minimal generalization of the familiar Bayesian framework. One could imagine further generalizations, but the framework suffices for our purposes.

To place this paper in the decision-theory literature, we note that our procedure (1)–(2) for assigning total weights of credence to subsets of outcomes can be viewed as a special case of the Dempster-Shafer theory of evidence (Dempster [9], Shafer [40]). In this theory, as here, the total belief in an event is the sum of the weights assigned to its subsets. The resulting *belief function* is also known as a *totally monotone capacity*. Moreover, the updating induced by our framework is equivalent to the Dempster-Shafer updating rule. This updating rule has been axiomatized by Gilboa and Schmeidler [17] in the context of Choquet expected utility maximization.[6]

Notice that we have restricted attention to deterministic hypotheses. One sees this in (1), where hypotheses are either clearly compatible or clearly incompatible with a given history. This is obviously restrictive—we are often interested in drawing inferences about theories that do not make sharp predictions. However, a model in which the implications of the evidence for various hypotheses is dichotomous simplifies the analysis by eliminating assessments as to which theories are more or less likely for a given history, in the process allowing us to focus attention on the resulting induction. Section 5.2 sketches the beginnings of a generalization to non-deterministic hypotheses.

Looking forward, it will be useful to have notation for the set of hypothe-

[5]In the process, the notion of compatibility needs to be adapted: whereas a single state ω is compatible with history h_t if $\omega \in [h_t]$, a (possibly multistate) hypothesis A is compatible with history h_t if $A \cap [h_t] \neq \varnothing$.

[6]No familiarity with this literature is required for the present paper. We came to work with the present model not because portions of it have antecedents in the literature, but because it captures other modes of reasoning while staying as close as possible to the familiar Bayesian model.

ses, *in a subset* $\mathcal{D} \subset \mathcal{A}$, that are relevant for prediction at history h_t:

$$\mathcal{D}(h_t) = \cup_{Y' \subsetneq Y} \{A \in \mathcal{D} \mid \varnothing \neq A \cap [h_t] \subset [h_t, Y']\}$$

Hence, $\mathcal{D}(h_t)$ is the set of hypotheses in $\mathcal{D}$ that have not been refuted by h_t and that could lend their weight to *some* non-tautological ($Y' \subsetneq Y$) prediction after history h_t, and $\phi(\mathcal{D}(h_t))$ will be the total weight of credence for these hypotheses.[7]

3 Special Cases

The unified framework is sufficiently general as to capture several standard models of inductive reasoning.

3.1 Bayesian Reasoning

We first show that our framework reduces to Bayesian reasoning if one restricts attention to hypotheses that consist of one state each.

Bayesian reasoning appeared explicitly in the writings of Bayes [2].[8] Beginning with the work of de Finetti and his followers, it has given rise to the Bayesian approach to statistics (see, for example, Lindley [28]). Relying on the axiomatic approach of Ramsey [34], de Finetti [7, 8], and Savage [37], it has grown to become the dominant approach in economic theory and in game theory. The Bayesian approach has also achieved great success in computer science and artificial intelligence, as in the context of Bayesian networks (Pearl [33]). Within the philosophy of science, notable proponents of the Bayesian approach include Carnap [4] and Jeffrey [23].

These manifestations of the Bayesian approach differ in several ways, such as the scope of the state space and the degree to which Bayesian beliefs are related to decision making, but they share two common ingredients: (i) uncertainty is always quantified probabilistically; and (ii) when new information is obtained, probabilistic beliefs are updated according to Bayes's rule.

[7] The restriction of attention to non-tautological hypotheses is introduced in order to render the numerical values of $\phi(\mathcal{D}(h_t))$, for different classes $\mathcal{D}$, more intuitive. Our main result also holds, and is in fact easier to prove, if one leaves in the set $\mathcal{D}(h_t)$ also hypotheses that are unrefuted and that do not restrict the prediction $y \in Y$ in any way.

[8] Precursors can be found in the early days of probability; see Bernoulli [3].

To embed Bayesian reasoning in our framework, define the set of *Bayesian hypotheses* to be

$$\mathcal{B} = \{\{\omega\} \mid \omega \in \Omega\} \subset \mathcal{A}. \tag{3}$$

Each of the Bayesian hypotheses thus fully specifies a single state of the world. In our price-of-oil example, a specific scenario might be that, at each t, demand for oil will increase, the level of conflict will not, and the price of oil will increase. This identifies a unique state ω with $\omega_X(t) = (1, 0)$ and $\omega_Y(t) = 1$ for all t, and the corresponding hypothesis is $A = \{\omega\}$.

A Bayesian agent will attach credence to no other hypotheses, i.e.,

$$\phi(A) = 0 \qquad \text{if} \qquad |A| > 1.$$

We can now state,

Observation 1 *Let p be a probability on Ω. There exists a model ϕ_p such that $\phi(\mathcal{A} \backslash \mathcal{B}) = 0$ and such that for every history h_t, there is a constant $\lambda > 0$ for which, for every $Y' \subset Y$*

$$p(y_t \in Y' \mid [h_t]) = \lambda \phi_p(\mathcal{A}(h_t, Y')).$$

This observation is verified by constructing the model $\phi_p(\{\omega\}) = p(\{\omega\})$, attaching to each hypothesis a weight of credence equal to the prior probability of the corresponding state. It is easy to verify that ϕ_p satisfies the equality of Observation 1 and that it is the unique such model.

Bayesian reasoning is thus a special case of our framework: every Bayesian belief can be simulated by a model ϕ, and Bayesian updating is imitated by our process of excluding refuted hypotheses. Apart from the normalization step, which guarantees that updated probabilities continue to sum up to 1 as hypotheses are deleted but has no effect on relative beliefs, Bayesian updating is nothing more than the exclusion of refuted hypotheses from further prediction.

In our example, a Bayesian hypothesis specifies the sign of changes in the level of demand, the level of conflict, and the price (all taking the values 0 or 1) in each of periods $t = 0, ..., T - 1$. The agent's prior distribution over these states is critical. Let us consider two possibilities. At one extreme is the *agnostic* Bayesian. This agent brings no knowledge to the problem, and hence places a uniform prior distribution over the set of states. Alternatively, we might consider a *confident* Bayesian who has clear ideas about which states

are likely. Suppose, for example, the agent believes that the market for oil is captured by a function $f(x^1, x^2)$, with the price of oil increasing in period t if and only if $f(x_t^1, x_t^2) > 0$. The agent's prior will accordingly attach positive probability only to states satisfying this relationship in each period, and will divide the prior probability among such states accordingly to how likely she finds the corresponding sequences of signs of changes in demand and conflict levels.

Given that our model captures Bayesian reasoning via an assumption that hypotheses contain only a single state, it is worth noting that an agent who assigns positive weight to non-Bayesian hypotheses (i.e., $\phi(\mathcal{A}\backslash\mathcal{B}) > 0$) will not be "Bayesian" by any common definition of the term. For example, suppose that $A = \{\omega_1, \omega_2\}$ and $\phi(A) = \delta > 0$. Such an agent can be viewed as arguing, "I think that one of ω_1 or ω_2 might occur, and I put a weight $\delta > 0$ on this hypothesis, but I cannot divide this weight between the two states." Intuitively, this abandons the Bayesian tenet of quantifying all uncertainty in terms of probabilities. Formally, if we use the resulting weight function to define a binary relation $\succsim$ over events, interpreted as "at least as likely as," we will find that such a relation will not satisfy de Finetti's [7, 8] cancellation axiom: it can be the case that, for two events, B, C, $B \succsim C$ but not $B\backslash C \succsim C\backslash B$. In addition, if we use the weight function to make decisions by maximization of a Choquet integral of a utility function, the maximization will fail to satisfy Savage's [37] "sure-thing principle"(axiom P2).[9] As a result, especially upon adding decisions to our model of beliefs (cf. Section 5.4), we have a converse to Observation 1: the decision maker will be Bayesian if and *only if* $\phi(\mathcal{A}\backslash\mathcal{B}) = 0$.

3.2 Case-Based Reasoning

Analogical reasoning was explicitly discussed by Hume [22], and received attention in the twentieth century in the guise of case-based reasoning (Riesbeck and Schank [36], Schank [38]), leading to the formal models and axiomatizations of Gilboa and Schmeidler [18, 19, 20].

[9] If positive weight is assigned to non-Bayesian hypotheses, one should specify how expected utility maximization is generalized to a theory of decision making where beliefs are given by a function ϕ that is not generally additive. A well-known such generalization is the maximization of a Choquet integral suggested by Schmeidler [39]. See Gilboa [14] for details and precise definitions. The axiomatic systems of de Finetti and Savage are also given in Kreps [26].

We consider here a very simple version in which case-based prediction is equivalent to kernel classification.[10] The agent has a similarity function over the characteristics,

$$s : X \times X \to \mathbb{R}_+,$$

and a memory decay factor $\beta \leq 1$. Given history $h_t = h_t(\omega) \in H_t$, a possible outcome $y \in Y$ is assigned the weight

$$S(h_t, y) = \sum_{i=0}^{t-1} \beta^{t-i} s(\omega_X(i), \omega_X(t)) \mathbf{1}_{\{\omega_Y(i)=y\}},$$

where $\mathbf{1}$ is the indicator function of the subscripted event. Hence, the agent may be described as if she considered past cases in the history h_t, chose all those that resulted in some period i with the outcome y, and added to the sum $S(h_t, y)$ the similarity of the respective characteristic $\omega_X(i)$ to the current characteristic $\omega_X(t)$. The resulting sums $S(h_t, y)$ can then be used to rank the possible outcomes y. If $\beta = 1$ and in addition the similarity function is constant, the resulting number $S(h_t, y)$ is proportional to the relative empirical frequency of y's in the history h_t.

To embed case-based reasoning in our framework, we first define case-based hypotheses as follows. For every $i < t \leq T - 1$, $x, z \in X$, let

$$A_{i,t,x,z} = \{\omega \in \Omega \mid \omega_X(i) = x, \omega_X(t) = z, \omega_Y(i) = \omega_Y(t)\}.$$

We can interpret this hypothesis as indicating that, *if* the input data in period i are given by x and are given in period t by z, *then* periods i and t will produce the same outcome (value of y). Notice that in contrast to the Bayesian hypotheses, a single case-based hypothesis consists of many states: $A_{i,t,x,z}$ does not restrict the values of $\omega_X(k)$ or $\omega_Y(k)$ for $k \neq i, t$.[11]

Let the set of all hypotheses of this type be denoted by

$$\mathcal{CB} = \{A_{i,t,x,z} \mid i < t \leq T, x, z \in X\} \subset \mathcal{A}. \tag{4}$$

For example, our oil-price predictor may focus only on the years in which demand and conflict had the same trends as in the current period, and make

[10] See Akaike [1] and Silverman [41].

[11] A case-based hypothesis thus contains a number of states (to be precise, $|Y|\left(|X||Y|\right)^{T-2}$) that is exponential in T. The hypothesis $A_{i,t,x,z}$ can only be refuted at periods i (if $\omega_X(i) \neq x$) or t (if $\omega_X(t) \neq z$ or $\omega_Y(i) \neq \omega_Y(t)$).

her prediction based on the prevalence of price increases in these periods. This would correspond to the similarity function

$$s\left(\left(x^1,x^2\right),\left(z^1,z^2\right)\right)=\begin{cases}1 & x^1=z^1,x^2=z^2\\ 0 & otherwise\end{cases}.$$

Alternatively, the agent may assign some weight also to past periods that resembled the current period only in one aspect, and use a similarity function such as

$$s\left(\left(x^1,x^2\right),\left(z^1,z^2\right)\right)=\begin{cases}1 & x^1=z^1,x^2=z^2\\ a & x^1=z^1,x^2\neq z^2\\ b & x^1\neq z^1,x^2=z^2\\ 0 & otherwise\end{cases}$$

for some $a,b\in(0,1)$.

We can now state:

Observation 2 *Let there be given $s:X\times X\to\mathbb{R}_+$ and $\beta\leq 1$. There exists a model $\phi_{s,\beta}$, such that $\phi(\mathcal{A}\backslash\mathcal{CB})=0$ and for every history h_t and every $y\in Y$,*

$$S(h_t,y)=\phi_{s,\beta}(\mathcal{A}(h_t,\{y\})).$$

This observation is verified by constructing the model

$$\phi_{s,\beta}(A_{i,t,x,z})=\beta^{(t-i)}s(x,z). \tag{5}$$

At history $h_t=h_t(\omega)$, only the hypotheses $\left\{A_{i,t,\omega_X(i),\omega_X(t)}\mid i<t\right\}$ yield predictions that are included in a singleton $\{y\}$. Hence, of the total number of case-based hypotheses,

$$|\mathcal{CB}|=|X|^2\binom{T}{2},$$

only t hypotheses will affect the prediction, corresponding to the t possible hypotheses of the form $A_{i,t,\omega_X(i),\omega_X(t)}$. These t hypotheses will be divided among the $|Y|$ possible values, each lending its weight to the outcome that occurred at the corresponding period i , $\omega_Y(i)$.

In general, we could define similarity relations based not only on single observations but also on sequences, or on other more general patterns of observations. Such higher-level analogies can also be captured as hypotheses

in our framework. For instance, the agent might find history h_t similar to history h_i for $i < t$, because in both of them the last k periods had the same observations. This can be reflected by hypotheses including states in which observations $(i-k+1), ..., i$ are identical to observations $(t-k+1), ..., t$, and so forth.

3.3 Rule-Based Reasoning

The earliest models of reasoning employing general rules date back to Greek philosophy and its study of logic, focusing on the process of deduction and the concept of proof. The rise of analytical philosophy, the philosophy of mathematics, and artificial intelligence greatly extended the scope of rule-based reasoning, including its use for modeling human thinking, as in the introduction of non-monotonic (McCarthy [30], McDermott and Doyle [31], Reiter [35]), probabilistic (Nilsson [32]), and a variety of other new logics.[12]

Various rules can be captured by assigning weights to appropriate hypotheses A in our framework. To consider an extreme example, the rule "the price of oil always rises" corresponds to the hypothesis

$$A = \{\omega \in \Omega \mid \omega_Y(t) = 1 \quad \forall t\}.$$

There are many states in this hypothesis, featuring different sequences of changes in the values of the level of demand and conflict.

Our framework can also encompass *association rules*, or rules that can be expressed as conditional statements. For example, consider the rule "if the level of conflict has risen, so will the price of oil." This rule can be described by

$$A = \{\omega \in \Omega \mid \omega_{X^2}(t) = 0 \quad or \quad \omega_Y(t) = 1 \quad \forall t\}. \tag{6}$$

(Recall that $\omega_{X^2}(t)$ indicates whether there was an increase in the index of conflict, and $\omega_Y(t)$ an increase in the price of oil. The rule "A implies B" is then read as "A is false, or B is true, or possibly both.")

A rule will be excluded from the summation defining $\phi(\mathcal{A}(h_t))$ as soon as a single counter-example is observed. Thus, if history h_t is such that for some $i < t$ we observed an increase in the level of conflict that was not followed by a rise in the price of oil, the hypothesis (6) will not be used for further analysis.

[12] See also Gardenfors [13] and Levi [27].

When an association rule is unrefuted, it may or may not affect predictions, depending on whether its antecedent holds. Specifically, if we consider a period t in which the level of conflict did *not* rise, the antecedent of rule A does not hold ($\omega_{X^2}(t) \neq 1$). This ensures that any value $\omega_Y(t)$ is compatible with A, and hence that the weight of the rule $\phi(A)$ will not be counted in the summation $\phi(\mathcal{A}(h_t, Y'))$ for any $Y' \subsetneq Y$. In general, if the antecedent of a rule is false, the rule becomes vacuously true and does not affect prediction. However, if (in this example) we do observe a rise in the level of conflict, $\omega_{X^2}(t) = 1$, the rule has bite (retaining the assumption that it is as yet unrefuted). Its weight of credence ϕ will be added to the prediction that the price of oil will rise, $\omega_Y(t) = 1$, but not to the prediction that it will not, $\omega_Y(t) = 0$.

Our framework also allows one to capture functional rules, stating that the value of y is a certain function f of the value of x, such as

$$A = \{\omega \in \Omega \mid \omega_Y(t) = f(\omega_X(t)) \quad \forall t\}.$$

Holland's [21] genetic algorithms employ additive aggregation over rules. This method addresses a classification problem where the value of y is to be determined by the values of $x = (x^1, ..., x^m)$, based on past observations of x and y. The algorithm maintains a list of association rules, each of which predicts the value of y according to values of some of the x^j's. For instance, one rule might read "if x^2 is 1 then y is 1" and another, "if x^3 is 1 and x^7 is 0 then y is 0." In each period, each rule has a weight that depends on its success in the past, its specificity (the number of x^j variables it involves) and so forth. The algorithm chooses a prediction y that is a maximizer of the total weight of the rules that predict this y and that apply to the case at hand.

The prediction part of genetic algorithms is therefore a special case of our framework, where the hypotheses are the association rules involved. However, in a genetic algorithm the set of rules and their associated weights do not remain constant, with rules instead being generated by a partly-random process, including crossover between "parent genes," mutations, and so forth.

There are many examples of rule-based reasoning that go beyond the simple cases we have just discussed. Indeed, *any* rule with a clear empirical meaning corresponds to a hypothesis A, which is its extension: the set of states of the world that are consistent with the rule.

3.4 Combined models

The previous subsections illustrate how our framework can capture each of the modes of reasoning separately. Its main strength, however, is in being able to smoothly combine such modes of reasoning, simply by considering models ϕ that assign positive weights to hypotheses of different types.

For example, consider an agent who attempts to reason about the world in a Bayesian way, to foresee all possible eventualities and assign probabilities to them. The agent has a probability p over the states of the world, Ω. However, in the back of her mind she also carries with her some general rules and analogies. Assume that she employs a model ϕ such that

$$\phi(\mathcal{B}) = 1 - \varepsilon$$

with weight allocated among the Bayesian hypotheses according to

$$\phi(\{\omega\}) = (1 - \varepsilon)\, p(\omega)$$

and the remaining weight ε is split among case-based and rule-based hypotheses.

If ε is small, the non-Bayesian hypotheses will play a relatively minor role is determining predictions, as long as history proceeds along a path that had a high a-priori probability p. However, suppose that the reasoner faces an eventuality, such as the September 11 attacks or the Lehman Brothers' collapse, that is surprising in the sense that the agent had assigned low or even zero probability p to this event. How will the agent make predictions? If she has assigned the event zero probability, Bayesian updating will not be well-defined. In this case, the non-Bayesian hypotheses, whose total weight is bounded by ε, may determine the agent's predictions. For example, in the face of the September 11 attack, the agent might discard Bayesian reasoning and resort to the general rule that "at the onset of war, the stock market plunges". Alternatively, the agent may resort to analogies, and predict the stock market's behavior based on past cases such as the attack on Pearl Harbor.

If the event in question had a nonzero but very small prior probability, non-Bayesian reasoning will again be relatively more important. For example, it is possible that the agent has conceived of the possibility of Lehman Brothers' collapse, but assigned a very small probability to this event. Once the event occurred, conditional probabilities are well-defined and can be used.

However, non-Bayesian hypotheses, which used to have a negligible effect on the reasoner's predictions, will now be much more prominent. This can be interpreted as if the reasoner has a certain degree of doubt about her own probabilistic assessments, captured by the weight $\varepsilon > 0$ put on non-Bayesian hypotheses. When a small probability event occurs, it as if the agent tells herself, "I do have my updated Bayesian beliefs, but I start doubting my probability assessments; after all, according to these very same assessment, it used to be very unlikely to find ourselves where we are. Hence, it might be a good idea to consider other modes of reasoning as well."

Our framework can thus describe the reasoning of agents who are mostly Bayesian most of the time. However, they have a certain degree of self-criticism that allows them to doubt their probabilistic assessments when they encounter surprises. Indeed, as we will see in the next section, it may not be easy for the reasoner to try to avoid surprises and at the same time to remain Bayesian.

4 Dynamics of Reasoning Methods

To study long-term effects, we consider a collection of frameworks, indexed by T, with the value of T growing arbitrarily large. The sets X and Y are assumed to remain the same for all T. In the framework with T periods there is a state space Ω_T, with a set of hypotheses $\mathcal{A}_T = 2^{\Omega_T}$. The agent uses a model ϕ_T. The sets of Bayesian ($\mathcal{B}_T$) and case-based ($\mathcal{CB}_T$) hypotheses for each T are defined as in (3) and (4).

4.1 Bayesian vs. non-Bayesian Reasoning

4.1.1 Assumptions

We provide conditions under which Bayesian reasoning fades into insignificance. We assume that at least some weight is placed on both Bayesian and case-based reasoning:

Assumption 1 *For some $\varepsilon > 0$, for every T,*

$$\phi_T(\mathcal{B}_T), \phi_T(\mathcal{CB}_T) > \varepsilon.$$

Assumption 1 means that the agent has a certain minimal degree of credence in both modes of reasoning, independent of T. Observe that there can

be many other types of hypotheses that get non-zero weight according to ϕ. We explain in Section 4.1.2 how this assumption could be reformulated to make no reference to case-based hypotheses.

Next, we can think of the agent as allocating the overall weight of credence in a top-down approach, first allocating weights to modes of reasoning, and then to specific hypotheses within each mode of reasoning. How should the weights be split within each set of hypotheses? We start with the weight of the Bayesian hypotheses, $\phi_T(\mathcal{B}_T)$. If the agent knows, or believes she knows something about the process she is about to observe, this knowledge should be reflected in her prior beliefs ϕ_T . In an extreme case, the agent might have $\phi_T(\{\omega\}) = 1$ for a particular ω. We are interested in the contrasting case of an agent who believes that she knows relatively little about the process she is observing. Such an agent cannot rule out *any* state and thus assigns a positive weight to each state ω.

How should these prior probabilities be chosen? A simple and common approach is to assume that the agent has a uniform prior over the state space, so that

$$\frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} = 1,$$

for any pair of states ω and ω', corresponding to the agnostic Bayesian of Section 3.1. This, however, is clearly restrictive. We seek a weaker assumption, requiring only that the probability assigned to any particular state cannot be too much smaller than that assigned to another state. Thus, one may assume that there exists $M > 0$ such that, for every $\omega, \omega' \in \Omega_T$ and every T,

$$\frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} < M. \tag{7}$$

We weaken this condition still further, allowing M to depend on T, and assume only that the ratio between the probabilities of two states cannot go to infinity (or zero) too fast as we consider ever-larger values of T. Formally,

Assumption 2 (Open-mindedness) *There exists a polynomial $P(T)$ such that, for every T and every two states $\omega, \omega' \in \Omega_T$,*

$$\frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} \leq P(T).$$

Assumption 2 allows for a more general class of beliefs than our first equal-probability or bounded-probability-ratios assumptions. In particular,

for each T, there exists $M = M_T$ satisfying (7), where M_T is allowed to tend to ∞ as $T \to \infty$. The assumption is, however, that this convergence is not too fast, in the sense that it is bounded by a polynomial in T.

Assumption 2 will typically be violated if, as often assumed in Bayesian models, the agent believes she faces successive independent and identically-distributed (iid) draws. For example, Assumption 2 will fail if, regardless of the levels of demand and conflict, the agent believes that the change in the price of oil is independently drawn to be positive ($\omega_Y(t) = 1$) in each period with probability $p > 0.5$. For an easy illustration of this failure, observe that the ratio of the probabilities of a string of T successive 1's and a string of T successive 0's is $(p/(1-p))^T$, and hence exponential in T. In the terms of Section 3.1, this is a confident Bayesian: because she believes that she knows quite a bit about the data generating process, she considers some states much more likely than others. Section 4.1.3 introduces and discusses generalizations of Assumption 2 that suffice for Theorem 1 and encompass such iid draws—our result thus continues to hold in such cases.[13]

We make an analogous assumption regarding the way that the weight of credence is distributed among the various case-based hypotheses. It would suffice for our result to impose a precise analog of Assumption 2, namely that there is a polynomial $Q(T)$ such that, for any T and any pair of case-based hypotheses $A_{i,t,x,z}$ and $A_{i',t',x'z'}$, we have

$$\frac{\phi_T(A_{i,t,x,z})}{\phi_T(A_{i',t',x',z'})} \leq Q(T). \tag{8}$$

However, suppose (analogously to (5)) that there exists a similarity function $s : X \times X \to \mathbb{R}_+$, a decay factor $\beta \in (0, 1]$, and, for every T, a constant $c_T > 0$ such that, for every $i < t < T$ and every $x, z \in X$,

$$\phi_T(A_{i,t,x,z}) = c_T \beta^{t-i} s(x, z). \tag{9}$$

In this case, the characteristics $x, z \in X$ determine the relative weights placed on the case-based hypotheses involving information of a given vintage (i.e., a given value of $t-i$), with $\beta \leq 1$ ensuring that older information is no more influential than more recent information. This formulation is rather natural, but it violates (8) if $\beta < 1$, as the relevance of older vintages then declines exponentially. Fortunately, there is an obvious and easily interpretable generalization of (8) that allows us to encompass (9).

[13] We do not introduce these generalizations here because they are less elegant and more cumbersome to interpret.

Assumption 3 *There exists a polynomial $Q(T)$ such that, (1) for every T, i, i', t, t', x, x' and z, z' with $t - i = t' - i'$,*

$$\frac{\phi_T(A_{i,t,x,z})}{\phi_T(A_{i',t',x',z'})} \leq Q(T) \tag{10}$$

and (2) for every T, $t < T$, $x, z \in X$ and $i < i' < t$,

$$\frac{\phi_T(A_{i,t,x,z})}{\phi_T(A_{i',t,x,z})} \leq Q(T). \tag{11}$$

Condition (10) stipulates that within a set of hypotheses based on similarities across a given time span (i.e., for which $t - i = t' - i'$), the agent's weights of credence cannot be too different. Condition (11) stipulates that when comparing similarities at a given period t, based on identical characteristics but different vintages, the older information cannot be considered too much *more* important than more recent information. Typically, we would expect older information to be *less* important and hence this constraint will be trivially satisfied.

4.1.2 Result

We now turn to the main result, stating that, under Assumptions (1)–(3), if the agent has a sufficiently long string of data, then she will discard Bayesian reasoning.[14]

Theorem 1 *Let Assumptions 1–3 hold. Then for every $\alpha, \delta > 0$ there exists T_0 such that, for every $T > \frac{1}{\alpha}T_0$, every $t \geq \alpha T$, and every history h_t,*

$$\frac{\phi_T\left(\mathcal{B}_T(h_t)\right)}{\phi_T\left(\mathcal{A}_T \backslash \mathcal{B}_T(h_t)\right)} < \delta.$$

In the most interesting case in which α and δ are taken to be quite small, Theorem 1 states that if the horizon is sufficiently long, then the agent will put virtually all of her weight on non-Bayesian (rather than on Bayesian) hypotheses for all but a small fraction of initial periods. Note

[14]Recall that for a subset of hypotheses $\mathcal{D} \subset \mathcal{A}$, the symbol $\phi(\mathcal{D}(h_t))$ is the total weight of the hypotheses in $\mathcal{D}$ that are effectively used for prediction at h_t. This is the weight of the hypotheses that have not yet been refuted, and that are not tautologically true at h_t.

that the only learning that is taking place in our process is the exclusion of refuted hypotheses. As mentioned in Section 2, this is the counterpart of Bayesian updating of probabilities. It is interesting that applying the logic of Bayesian updating to all hypotheses, and thereby also to the choice of reasoning methods, favors non-Bayesian reasoning.

Importantly, this theorem says nothing about what would be correct or optimal learning. It does not compare the agent's predictions to any external standard or truth. Rather, it is a description of the evolution of the agent's reasoning.

The Bayesian part of the agent's beliefs converges to the truth at an exponential rate as evidence is accumulated (that is, as t grows)—within this class of hypotheses, the probability of the true state *relative to* the probability of all unrefuted states grows exponentially with t.[15] This increase of the posterior probability of the true state does not result from any change in the prior probability, but from the exclusion of falsified states. In other words, the conditional probability of the true state increases at an exponential rate because its denominator, given by the total probability of all unrefuted states, decreases at an exponential rate. But as we now explain, this is precisely the reason that the weight of the entire class of Bayesian hypotheses tapers off and leaves the stage to others, such as the case-based hypotheses. The very mechanism that makes the posterior probability of truth grow makes the weight of Bayesian reasoning diminish.

In particular, as periods pass and data accumulate, the total weight of the case-based hypotheses increases as compared to that of the Bayesian ones. The key observation is that for long horizons, there are many more Bayesian than case-based hypotheses, with the number of the former increasing exponentially in T while the latter increase polynomially (in fact quadratically) in T. In addition, as t grows (for fixed T), the number of hypotheses that remain unrefuted by the history h_t becomes an exponentially small fraction of the size of $\mathcal{B}_T$, and thus (given open-mindedness) their cumulative weight tends to zero at an exponential rate. In contrast, the fraction of case-based hypotheses that are unrefuted by the history h_t and that make nontrivial period-t predictions increases in t. The *relative* weight placed on Bayesian hypotheses thus declines.

[15] As a result, there are priors consistent with Assumption 2 for which the relative probability attached to the true state, within the class of Bayesian hypotheses, gets arbitrarily close to 1 as data accumulate.

It follows that a similar result would hold if we were to replace the class of case-based hypotheses with any other class of hypotheses that grows polynomially in T and that provides some non-tautological prediction for each h_t. Therefore, we do not view this theorem as extolling the virtues of case-based reasoning. Case-based reasoning is simply a familiar example of a mode of reasoning with the requisite properties. The theorem points to general circumstances under which Bayesian reasoning will tend to wither away, for reasons unrelated to bounded rationality or to cognitive or computational limitations.

We can gain some insight into the potential difficulties with Bayesian reasoning by linking the evolution of the weights attached to Bayesian and case-based hypotheses to the structure of the hypotheses. A case-based hypothesis contains (exponentially in T) many states, each of which by itself corresponds to a Bayesian hypothesis. As t increases, increasing numbers of these Bayesian hypotheses will be refuted and their weight will melt away from the total weight of Bayesian hypothesis. By contrast, a smaller proportion of case-based hypotheses are refuted at each stage. As a result, the weight commanded by the case-based hypothesis grows relative to that of its constituent Bayesian hypotheses. It then seems as if the whole (the case-based hypothesis) retains more weight than the sum of its parts (the Bayesian hypotheses). The reason is that case-based hypothesis need not divide its weight among its states. Hence, it need not say too much, and is consequently not so vulnerable to refutation.

To illustrate this, consider our price-of-oil example, and, let us develop intuition by thinking of each hypothesis as the theory proposed by a particular expert. Let us begin fifty years back. The year is 1960 and different experts are asked about the evolution of the price of oil over the next sixty years. A Bayesian expert provides a hypothesis that consists of a single state. There are many Bayesian experts, and each gets a small a priori weight. To make the calculations transparent, assume that a total weight of $(1-\varepsilon)$ is divided equally among the 8^{60} hypotheses.

We are now fifty years later, in the year 2010, we have just observed this year's demand and conflict indicators, and we are trying to predict the remaining ten years. Whatever was the history h_{50}, there are 2×8^9 states still consistent with it. That is, the proportion of Bayesian hypotheses that are still in the game is

$$\frac{2 \times 8^9}{8^{60}} < 8^{-50}.$$

Under a uniform distribution assumption, the total weight of these hypotheses is bounded by $8^{-50}(1-\varepsilon)$. We emphasize, however, that the uniform distribution is not essential here. Instead, as Assumption 2 demands, what is important is that a set of hypotheses whose size diminishes exponentially fast (in t) should also have a total weight that diminishes exponentially fast.

Thus, the vast majority of the Bayesian experts will have been proven wrong and only a few will still be in the game, resulting in an overall low weight for the Bayesian experts as a group (even though the expert who happens to be right will have large *relative* standing within this group). In contrast, consider a case-based expert who says "the price of oil in year 2010 will be similar to that in year 1991." This hypothesis is clearly not Bayesian, as it says nothing about the years 1960, 1961,...,1990, 1992,...,2009. Saying nothing about these years, the hypothesis is not risking being refuted by the respective observations. Moreover, each year there are more such unrefuted, relevant case-based hypotheses, as the number of previous years (and hence the number of cases) grows. As a result, many more case-based hypotheses remain unrefuted by the year 2010, causing case-based hypotheses to dominate the agent's reasoning.

Suppose the agent demands that experts completely specify their hypotheses, that is, that all be Bayesian. Hence, the non-Bayesian (case-based) expert who predicted that the price of oil in 2010 would be similar to 1991 is asked to divide the weight of her hypothesis among the exponentially many states that constitute her hypothesis. This expert might reasonably reply, "I do not have any prediction about 1960 or about 1971. All I said was that there will be similar outcomes in 1991 and in 2010." Insisting that the expert further specify the hypothesis may be stretching the limit of her expertise. And if later it is found out that her predictions for the years 1960 or 1971 were falsified, it would be wrong to penalize her original prediction, which did not purport to say anything about these observations.

Put differently, the Bayesian approach is not flexible enough to allow the experts to say "I do not know." It requires that they quantify their beliefs about all questions. Hence, the Bayesian approach does not allow us to distinguish among experts according to the accuracy of their self-assessment. An expert who knows that she does not know certain probabilities and an expert who wrongly believes that she does know them may end up subscrib-

ing to the same assessments.[16] In contrast, other approaches allow for a "don't know" answer, and thus can indirectly give experts credit for knowing the limitations of their knowledge. A non-Bayesian expert may avoid refutation either by making correct predictions, or by knowing when to remain silent. The inclusion of non-Bayesian hypotheses therefore allows the agent to judge experts not only by their specific knowledge, but also by their meta-knowledge, namely, the knowledge of what they know and what they do not know.

4.1.3 Generalizations

Section 4.1 mentioned that Assumption 2 typically does not hold if the agent believes she observes an independent and identically distributed (iid) process, as in the example of a sequence of Bernoulli random variables with $p \neq 0.5$. On the one hand, we would have been neither surprised nor distressed if the result of Theorem 1 failed for Bayesians who believe they face iid processes. The message of Theorem 1 is that Bayesian reasoning will not persist unless it is based on sufficiently precise prior information. We view the iid case as one in which the agent believes she has quite precise prior information, in the sense that she believes she knows the data generating process up to the specification of a single parameter. We find the theorem of interest because there are interesting induction problems where such precise prior information is *not* available.

It turns out, however, that the conclusion of Theorem 1 continues to hold for the case of a Bayesian who believes she faces a sequence of independent Bernoulli random variables with $p \neq 0.5$. Such a process ensures that the maximal weight attached to any single Bayesian hypothesis is shared by exponentially many such hypotheses, which implies that this weight becomes exponentially small as T increases, and this suffices for the result. In particular, of T observations, one would expect to find roughly pT realizations $y = 1$. Assuming that pT is an integer, the most likely states are those with precisely pT realizations $y = 1$. However, there are $\binom{T}{pT}$ such states, a number that grows exponentially in T. Even though some other hypotheses

[16] For example, the Bayesian approach does not allow us to distinguish a prediction of 50/50 (perhaps of the probability of rain tomorrow) that is based on complete ignorance from a prediction based on precise knowledge of the underlying process. This critique of the Bayesian approach was explicitly stated in Ellsberg [10] and Schmeidler [39] and can be traced back to Knight [25].

are very much less likely, the existence of exponentially many hypotheses of comparable and relatively large weights suffices for the result.

This suggests a generalization of Assumption 2 that would suffice for our theorem to hold: assume that there exist a polynomial $P(T)$ and, for every T, a subset $A_T \subset \Omega_T$, such that $|A_T|$ grows exponentially in T, and, for every T and every two states $\omega \in \Omega_T$, $\omega' \in A_T$,

$$\frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} \leq P(T).$$

This covers common applications of Bayesian reasoning, most notably applications to iid random variables.[17]

While this generalization is both straightforward and brings many familiar Bayesian problems within the purview of Theorem 1, it also raises some critical questions. Essentially, our iid Bayesian is attempting to predict the precise sequence of observations. Why would she do this, instead of restricting herself to predicting sufficient statistics, such as the average of the random variables? Some care would be needed here, because a precise prediction of the sequence of averages is equivalent to a precise prediction of the random variables themselves. We might try to avoid this by allowing the agent to predict only ranges of the relevant statistics, but in the current framework this will again give rise to exponentially-many most likely hypotheses and hence to the result of Theorem 1. The derivation of conditions that guarantee the survival of Bayesian reasoning with iid random variables is an interesting and important question. We think this question is best pursued in a generalization of our current framework to encompass probabilistic hypotheses (cf. Section 5.2).

4.1.4 When will Bayesianism Prevail?

Theorem 1 shows that in some circumstances, Bayesian reasoning is guaranteed to die out and leave the stage to others. There are also circumstances under which Bayesian reasoning will remain useful and even dominant. To begin with a trivial example, consider an agent who is a devout Bayesian, satisfying

$$\phi_T(\mathcal{B}_T) = 1$$

[17] Generalizing further, it would suffice for a somewhat weaker version of Theorem 1 that the weight attached to the most likely Bayesian hypothesis declines exponentially in T.

for all T (and hence failing Assumption 1). Such an agent will obviously remain Bayesian in the face of whatever evidence she gathers. However, should the agent allow for the smallest doubt and assign a positive weight either to a case-based or to a rule-based mode of reasoning, then the Bayesian way of thinking will be driven out by its competitors. Interpreting the weights as subjective probabilities regarding the theory that actually governs the data generating process, it suffices that a very small probability is assigned to the non-Bayesian ways of thinking, for our generalized Bayesian updating to shrink the weight put on the Bayesian approach.

Possible violations of Assumption 2 provide more useful insights into the realms in which Bayesian reasoning will prevail. First, Assumption 2 is obviously violated if the agent believes that she nearly knows the true state of the world, say, if for some ω_T, $\phi_T(\{\omega_T\}) = 1 - \varepsilon$ for all T. If, on top of this, the agent is also correct in her focus on state ω_T, then (that is, at state ω_T) the posterior probability attached to Bayesian hypotheses will never dip below $1 - \varepsilon$. In other words, if the agent believes she knows the truth, and *happens to be right*, her Bayesian beliefs will remain dominant.

A slightly less trivial example is the following. Consider, for concreteness, only Bayesian and case-based reasoning. For simplicity, let $X = \{0\}$ and $Y = \{0, 1\}$, so that all periods have the same observable features, and they only differ in the binary variable the agent is trying to predict. Suppose the agent believes that she observes a cyclical process. Formally, for $1 \leq k \leq T$, let $\omega^k \in \Omega_T$ be defined by

$$\omega_Y^k(t) = \begin{cases} 0 & 2mk \leq t < (2m+1)k \quad m = 0, 1, 2, ... \\ 1 & (2m+1)k \leq t < (2m+2)k \quad m = 0, 1, 2, ... \end{cases}.$$

Thus, for $k = 1$ the process is 01010101..., for $k = 2$ it is 001100110011... and so forth.

Let the agent's beliefs satisfy

$$\phi_T(\{\omega^k\}) = \frac{1 - \varepsilon}{2^k}$$

and

$$\phi_T(\{\omega\}) = 0$$

for every $\omega \notin \left\{\omega^k \,|1 \leq k \leq T\right\}$. Thus, the agent splits all the weight of the Bayesian hypotheses among the k hypotheses $\left\{\omega^k\right\}$ and leaves no weight to

the other Bayesian beliefs.[18] The remaining weight, $\hat{\varepsilon} = \varepsilon + \frac{1-\varepsilon}{2^T}$, is split equally among the case-based hypotheses.

Next suppose that the agent is right in her belief that the process is cyclical (starting with a sequence of 0's). Thus, the data generating process chooses one of the states ω^k. At this state, once we get to period $t = k$, all the Bayesian hypotheses $\left\{\omega^{k'}\right\}$ for $k' \neq k$ are refuted. In contrast, the hypothesis $\left\{\omega^k\right\}$ is not refuted at any t. Consequently, at ω^k, for every $t \geq k$, the total weight of the Bayesian hypotheses remains $\frac{1-\varepsilon}{2^k}$. In contrast, the total weight of the case-based hypotheses can never exceed $\hat{\varepsilon}$, resulting in the Bayesian mode of reasoning remaining the dominant one (for small ε and large T). Clearly, this will only be true at the states $\left\{\omega^k\right\}$. At other states the converse result holds, because all Bayesian hypotheses will be refuted and case-based reasoning will be the only remaining mode of reasoning.

Two main assumptions are thus needed for the success of the Bayesian approach. First, the agent has to believe that some states are very much more likely than others.[19] Second, she has to be right. If the agent knows the data generating process up to certain parameters (such as the cycle length, k, in the last example), then Bayesian beliefs allow learning and need not be driven out by other forms of reasoning. In these circumstances, especially if the parameters can assume only finitely many values or belong to a compact space, it makes sense to think of the agent's prior beliefs as being specified over the set of parameters (rather than over the state space Ω_T) and to update that prior as observations are gathered. This type of Bayesian reasoning is successful because the set of parameters does not increase with the number of observations, even though the set of states does. Put differently, the agent is not learning the state space Ω_T, which grows with T, but the parameter space, which is fixed. The Bayesian approach will then be quite successful.

In contrast, assume that the agent observes a process about which nothing is known a priori. She may be interested in the price of oil, the rise and fall of economic powers, or the eruption of wars. In these cases there is no claim that the data generating process is known up to a finite set of parameters,

[18] Observe that these Bayesian beliefs can also be described as rule-based beliefs. We suspect that this is not a coincidence. When Bayesian beliefs violate Assumption 2, it is likely to be the case that they reflect some knowledge about the data generating process, which can also be viewed as believing in a class of rules.

[19] Notice that Assumption 2 already allows large differences in the prior probabilities attached to various states, and Bayesian reasoning can survive only if the agent is yet *more* convinced of the differences between various states.

and hence the agent has to form her beliefs over the entire state space rather than over a parameter space. Moreover, the size of the state space increases at an exponential rate with the horizon T. It then seems a priori harder to learn the process. Correspondingly, we find Assumption 2 rather natural for such examples. In fact, one can argue that it is irrational to violate it, as such a violation suggests that the agent believes she knows about the process more than she actually does.

4.2 Case-Based vs. Rule-Based Reasoning

We devote this subsection to another application of the framework, dealing with the dynamics of case-based versus rule-based reasoning. We provide a simple example that shows how the relative weights of these modes of reasoning change endogenously.

Assume for convenience that there are no predicting variables, or equivalently, that the value of x is constant: $|X| = 1$. Let $Y = \{0, 1\}$ and assume that y_t are iid, where $y_t = 1$ with probability p.

Consider the set of rules,

$$\mathcal{RB}_T = \{R_{i,y} \mid i < T, y \in Y\},$$

where

$$R_{i,y} = \{\omega \in \Omega \mid \omega_Y(t) = y \quad \forall t \geq i\}$$

for $i \geq 0$ and $y \in Y$. Hence, each rule is identified by a given period i and outcome y, and predicts that from period i on, only outcome y will be observed. There are $2T$ hypotheses in $\mathcal{RB}_T$.

Because there are no x values to consider, the case-based hypotheses are simply

$$A_{i,t} = \{\omega \in \Omega \mid \omega_Y(i) = \omega_Y(t)\},$$

and the set of all case-based hypotheses is

$$\mathcal{CB}_T = \{A_{i,t} \mid i < t \leq T\},$$

containing $T(T-1)/2$ hypotheses.

Assume that, for $0 < c < 1$

$$\begin{aligned} \phi_T(\mathcal{CB}_T) &= c \\ \phi_T(\mathcal{RB}_T) &= 1 - c, \end{aligned}$$

and, for simplicity, that the weights within each class of hypotheses are uniformly distributed. Thus, for $i \geq 0$ and $y \in \{0, 1\}$,

$$\phi_T(R_{i,y}) = \frac{1-c}{2T},$$

and for $i < t \leq T$,

$$\phi_T(A_{i,t}) = \frac{2c}{T(T-1)}.$$

Next, let us consider histories h_t ending with k $(0 \leq k \leq t)$ 1's, that is, $y_i = 1$ for $t-k \leq i < t$, but $y_{t-k-1} = 0$ (or $k = t$) . For each $t-k \leq i \leq t$, the rule $R_{i,1}$ is unrefuted and non-tautological at h_t and predicts $y_t = 1$. There is one more rule that is unrefuted and non-tautological at h_t, and this is $R_{t,0}$, which predicts 0. Overall the rule-based prediction contributes $\frac{(k+1)(1-c)}{2T}$ to the prediction 1, and $\frac{1-c}{2T}$ to the prediction 0. The case-based prediction splits the weight $\frac{2tc}{T(T-1)}$ between 0 and 1 proportionally to the average

$$\bar{y}_{t-1} = \frac{1}{t}\sum_{i=0}^{t-1} y_i.$$

Assume that T is large. For small values of t, the overall weight of the case-based prediction is $O(T^{-2})$ and it therefore does not significantly change the rule-based prediction. Thus, even if k is small, the agent will predict $y_t = 1$. This phenomenon means that the agent is a little too quick to find trends in the data. A few observations of the value 1 suffice for her to theorize that "from now on, we'll observe only 1." We may view this phenomenon as a type of overfitting: the agent's function ϕ assigns a high weight to a theory that matches the data perfectly, even though this theory only deals with the most recent observations.

Next consider large values of t, say, $t = \alpha T$ for $\alpha \in (0, 1)$. The total weight of the case-based hypotheses is now

$$\phi_T\left(\mathcal{CB}_T\left(h_t\right)\right) = \frac{2tc}{T(T-1)} = \frac{2\alpha c}{(T-1)},$$

and it is of the same order of magnitude as the total weight of the rule-based predictions,

$$\phi_T\left(\mathcal{RB}_T\left(h_t\right)\right) = \frac{(k+2)(1-c)}{2T}.$$

The ratio of the two is

$$\frac{\phi_T\left(\mathcal{CB}_T\left(h_t\right)\right)}{\phi_T\left(\mathcal{RB}_T\left(h_t\right)\right)} = \frac{\frac{2\alpha c}{(T-1)}}{\frac{(k+2)(1-c)}{2T}} = 4\alpha \times \frac{T}{T-1} \times \frac{c}{1-c} \times \frac{1}{k+2}.$$

It follows that for low values of c, reflecting a tendency to theorize, the agent will tend to over-generalize and find patterns even in data that are in fact random. In contrast, high values of c, associated with a tendency to rely on experience, will reduce the chance of overfitting the data and over-generalizing trends, at the cost of ignoring trends when they actually exist.

Intermediate values of c would result in rule-based reasoning being dominant for large k (relative to c and α), and case-based reasoning taking over when k is small. In other words, when recent history is suggestive of a simple rule (a large number of observations of 0 or of 1), the agent adopts the rule "recent observations will continue forever." When recent history is more spotty, and no simple rule explains it, the agent assigns less weight to rule-based reasoning and resorts to case-based reasoning, which in this case means reliance on past frequencies. Since, for every k, there is a positive probability to observe a run of k 0's or k 1's, for a large T we should expect to find periods in which history suggests rules, followed by periods in which no rule seems to explain the data. Therefore, it should be expected that from time to time there will emerge a theory that is accepted by most agents, and at some point it will collapse. When it does collapse, confusion may lead agents to adopt less theoretical, more case-based methods, until the data seem to suggest a new theory, and so forth. In other words, even if the data are completely random, it should be expected that theories would rise and fall every so often, with case-based reasoning being more prominent between regimes of different theories.

Observe that the balance of weights between the two modes of reasoning is driven by the success of rule-based reasoning. This reflects the intuition that people would like to understand the process they observe, and that such "understanding" means a simple, concise theory that explains the data. If such a theory exists, agents will tend to prefer it over case-based reasoning. But when all simple theories are refuted, agents will resort to case-based reasoning. Theories or rules are exciting when they succeed, but, being ambitious, they can also fail. Cases, by contrast, are no more than an amalgamation of data, and thus they do not provide any deep insights or a sensation of "understanding."On the bright side, they can never be refuted. They are always

there, waiting faithfully for the agent, who would devote more attention to them when her heroic attempts to understand the process fail.

5 Discussion

5.1 Methods for Generating Hypotheses

In many examples ranging from scientific to everyday reasoning, it may be more realistic to put weight ϕ not on specific hypotheses A, but on methods or algorithms that generate them. For example, linear regression is one such method. When deciding how much faith to put in the prediction generated by the OLS method, it seems more plausible that agents put weight on "whatever the OLS method prediction came out to be" rather than on a specific equation such as "$y_t = 0.3 + 5.47x_t$."

One simple way to capture such reasoning is to allow the carriers of weight of credence, that is, the argument of ϕ, to be sets of hypotheses, with the understanding that within each set a most successful hypothesis is selected for prediction, and that the degree of success of the set is judged by the accuracy of this most successful hypothesis. The following example illustrates.

Suppose that the agent is faced with a sequence of datasets. In each dataset there are many consecutive observations, indicating whether a comet has appeared (1) or not (0). Different datasets refer to potentially different comets.

Now assume that the agent considers the general notion that comets appear in a cyclical fashion. That is, each dataset would look like

$$0, 0, ..., 0, 1, 0, 0, ..., 0, 1, ...$$

where a single 1 appears after k 0's precisely. However, k may vary from one dataset to the next. In this case, the general notion or "paradigm" that comets have a cyclical behavior can be modeled by a set of hypotheses—all hypotheses that predict cycles, parametrized by k. If many comets have been observed to appear according to a cycle, the general method, suggesting "find the best cyclical theory that explains the observations" will gain much support, and will likely be used in the future. Observe that the method may gain credence even though the particular hypotheses it generates differ from one dataset to the next.

5.2 Probabilistic Hypotheses

An important next step is to extend this framework to probabilistic hypotheses. Hypotheses would then be represented by probability distributions rather than by sets of states. The Bayesian hypotheses in such an extension are straightforward, and consist of probability distributions over states. Each such distribution f has an a priori weight $\phi(\{f\})$. If the support of ϕ is contained within the set of Bayesian hypotheses, then ϕ is simply the Bayesian prior. Given a history h_t, the hypothesis f is no longer classified dichotomously into "consistent with h_t" or "inconsistent with h_t." Rather, it is continuously ranked in $[0, 1]$ according to the probability of history h_t given theory f, that is, according to the theory's likelihood function at h_t. Multiplying the likelihood function by the a-priori weight $\phi(\{f\})$ leads to a natural measure of the belief in theory f following history h_t. Indeed, this is, up to renormalization, precisely the result of a Bayesian update over the Bayesian hypotheses.

The specification of non-Bayesian hypotheses is less clear. Should these be formulated as sets of distributions over states, or as distributions over sets of states, some combination of these generalizations, or something else? Finding such an appropriate generalization is a topic for further research.

5.3 Single-Hypothesis Predictions

This paper is concerned with reasoning that takes many hypotheses into account and aggregates their predictions. Alternatively, we may consider reasoning modes that focus on a most preferred hypothesis (among the unrefuted ones) and make predictions based on it alone. For example, if we select the simplest theory that is consistent with the data, we obtain Wittgenstein's [43] definition of induction.[20] If, by contrast, we apply this method to case-based hypotheses, we end up with nearest-neighbor approaches (see Cover and Hart [6] and Fix and Hodges [11, 12]) rather than with the case-based aggregation discussed here.

[20] See Solomonoff [42], who suggested to couple this preference for simplicty with Kolmogorov complexity measure to yield a theory of philosophy of science. Gilboa and Samuelson [15] discuss the optimal selection of the preference relation over theories in this context.

5.4 Decision Theory

The present paper deals with prediction. In order to explore its implications to decision making, the framework needs to incorporate acts and payoffs, and to specify the interaction between the agent's choices and the underlying process.

One simple possibility is to assume that the agent makes one choice of an act (or a strategy) at the outset, then history unfolds, nature determines the state of the world, and the agent's utility is determined by the resulting outcome. In this case, each act f associates outcomes with states ω as in a standard Savage model.

Because a model ϕ assigns non-negative weights to subsets of a state space, Ω, it defines a totally monotone capacity:

$$v(A) = \sum_{B \subset A} \phi(B).$$

It is therefore natural to define the agent's preferences by the Choquet integral (Choquet [5]) of her utility, as axiomatized by Schmeidler [39]. Observe that the non-additivity of the capacity results from the fact that a weight given to a certain hypothesis cannot always be divided among the individual states in it. When combined with the Choquet integral, this partial information results in ambiguity-averse behavior.

An additional source of ambiguity may be the model ϕ. While we assumed that the agent has a single such model, it may be more realistic to allow a certain degree of model uncertainty. In this case, one may consider a set of models, Φ. Decisions may then be taken in ways that are analogous to the multiple priors decision theories, such as maxmin expected utility (Gilboa and Schmeidler [16]), variational preferences (Maccheroni, Marinacci, Rustichini [29]), smooth preferences (Klibanoff, Marinacci, and Mukerji [24]), and so forth. Such a model can capture two types of learning: the generalized Bayesian learning that consists of excluding refuted hypotheses, as well as the learning that consists of shrinking the set Φ, inspired by classical statistics methods.

6 Appendix: Proof of Theorem 1

Let there be given $\alpha, \delta > 0$. We need to show that there exists T_0 such that, for every $T > \frac{1}{\alpha}T_0$, every $t \geq \alpha T$, and every history h_t,

$$\frac{\phi_T\left(\mathcal{B}_T(h_t)\right)}{\phi_T\left(\mathcal{CB}_T(h_t)\right)} < \delta.$$

We first bound the numerator from above, using Assumption 2. Then we bound the denominator from below, using Assumption 3.

We start by showing that, because the ratio of weights assigned to specific states (hypotheses in $\mathcal{B}_T$) is bounded by a polynomial, the weight of each particular state is bounded by the polynomial divided by an exponential function of T.

Consider a state ω. If $\phi_T(\{\omega\}) > \eta$, then, since for every $A, B \in \mathcal{B}_T$, $\phi_T(A) \leq P(T)\phi_T(B)$, for every ω',

$$\phi_T(\{\omega'\}) \geq \frac{\phi_T(\{\omega\})}{P(T)} > \frac{\eta}{P(T)}$$

Observe that $|\Omega| = d^T$ for $d = |X||Y| > 1$. Hence

$$\phi_T(\mathcal{B}_T) > \frac{d^T\eta}{P(T)}$$

and $\phi_T(\mathcal{B}_T) < 1$ implies

$$\eta < \frac{P(T)}{d^T}$$

Since this is true for every η such that $\eta < \phi_T(\{\omega\})$, we conclude that

$$\phi_T(\{\omega\}) \leq \frac{P(T)}{d^T}.$$

Finally, recall that ω was arbitrary, hence this inequality holds for every ω.

Next, we wish to show that the weight of all the states that are consistent with a given history h_t has to be relatively small. Observe that the number of states that are consistent with h_t is exponential in $(T - t)$. Indeed, if t were fixed, their total weight need not converge to zero. However, we assume that t is at least a fixed proportion, α, of T. This implies that the total weight of

all states consistent with h_t decreases exponentially with T. Specifically, for every h_t,

$$|\mathcal{B}_T(h_t)| = \frac{d^{T-t}}{|X|} \leq d^{T-t}$$

and it follows that

$$\phi_T\left(\mathcal{B}_T(h_t)\right) \leq d^{T-t}\frac{P(T)}{d^T} = \frac{P(T)}{d^t}$$

and since $t \geq \alpha T$,

$$\phi_T\left(\mathcal{B}_T(h_t)\right) \leq \frac{P(T)}{d^{\alpha T}} = \frac{P(T)}{\left(d^\alpha\right)^T} \tag{13}$$

where $\alpha > 0$ and thus $d^\alpha > 1$.

We now turn to discuss the weight of the case-based hypotheses. We wish to show that this weight cannot be too small. Consider a hypothesis $A_{(t-1),t,x,z} \in \mathcal{CB}_T$ and assume that $\phi_T\left(A_{(t-1),t,x,z}\right) < \xi$. By (10) (of Assumption 3) we have that, for all t', x', z'

$$\phi_T\left(A_{(t'-1),t',x',z'}\right) < \xi Q(T).$$

By (11) (of that Assumption), we know that for all $i < t' < T$, and all x', z',

$$\phi_T\left(A_{i,t',x',z'}\right) < \phi_T\left(A_{(t'-1),t',x',z'}\right) Q(T) < \xi\left[Q(T)\right]^2.$$

As the overall number of case-based hypotheses is $|X|^2 \begin{pmatrix} T \\ 2 \end{pmatrix}$, we obtain the bound

$$\phi_T\left(\mathcal{CB}_T\right) < \xi\left[Q(T)\right]^2 |X|^2 \begin{pmatrix} T \\ 2 \end{pmatrix}.$$

Define

$$R(T) = \left[Q(T)\right]^2 |X|^2 \begin{pmatrix} T \\ 2 \end{pmatrix}$$

and observe that it is a polynomial in T.

Thus, we have

$$\xi > \frac{\phi_T\left(\mathcal{CB}_T\right)}{R(T)}.$$

Since this holds for any ξ such that $\xi > \phi_T\left(A_{(t-1),t,x,z}\right)$, it has to be the case that

$$\phi_T\left(A_{(t-1),t,x,z}\right) \geq \frac{\phi_T\left(\mathcal{CB}_T\right)}{R(T)}$$

and $\phi_T\left(\mathcal{CB}_T\right) > \varepsilon$ implies

$$\phi_T\left(A_{(t-1),t,x,z}\right) \geq \frac{\phi_T\left(\mathcal{CB}_T\right)}{R(T)} > \frac{\varepsilon}{R(T)}.$$

We observe that at h_t there are precisely t case-based hypotheses that are unrefuted and non-tautological, and among them there is one of the type $A_{(t-1)t,x,z}$ (that is, the one defined by $x = \omega_X(t-1)$ and $z = \omega_X(t)$). It follows that

$$\phi_T\left(\mathcal{CB}_T(h_t)\right) \geq \phi_T\left(A_{(t-1),t,x,z}\right) > \frac{\varepsilon}{R(T)}. \tag{14}$$

Since (13) yields

$$\phi_T\left(\mathcal{B}_T(h_t)\right) \leq \frac{P(T)}{\left(d^\alpha\right)^T}$$

and (14) yields

$$\frac{1}{\phi_T\left(\mathcal{CB}_T(h_t)\right)} < \frac{R(T)}{\varepsilon},$$

we can conclude that

$$\frac{\phi_T\left(\mathcal{B}_T(h_t)\right)}{\phi_T\left(\mathcal{CB}_T(h_t)\right)} < \frac{P(T)R(T)}{\varepsilon\left(d^\alpha\right)^T}.$$

As $P(T)$ and $R(T)$ are polynomials in T and the denominator is an exponential function in T with base $d^\alpha > 1$, for a large enough T_0 and hence T, this ratio will be below the specified δ. ■

References

[1] Hirotugu Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6(2): 127–132, 1954.

[2] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418, 1763. Communicated by Mr. Price.

[3] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713.

[4] Rudolf Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.

[5] Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5 (Grenoble): 131–295, 1953–54.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27, 1967.

[7] Bruno de Finetti. Sul Significato Soggettivo della Probabilità. *Fundamenta Mathematicae*, 17: 298-329, 1931.

[8] Bruno de Finetti. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincare*, 7(1): 1–68, 1937.

[9] Arthur. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2): 325–339, 1967.

[10] Daniel Ellsberg. Risk, ambiguity and the Savage axioms. *Quartely Journal of Economics*, 75(4): 643-669, 1961.

[11] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical report 4, project number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[12] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discrimination: Small sample performance. Report A193008, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

[13] Peter Gärdenfors. Induction, conceptual spaces and AI. *Philosophy of Science*, 57(1): 78–95, 1990.

[14] Itzhak Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, Cambridge, 2009.

[15] Itzhak Gilboa and Larry Samuelson. Subjectivity in inductive inference. Cowles Foundation Discussion Paper 1725, Tel Aviv University and Yale University, 2009.

[16] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18: 141–153, 1989.

[17] Itzhak Gilboa and David Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59(1): 33–49, 1993.

[18] Itzhak Gilboa and David Schmeidler. Case-based decision theory. *Quarterly Journal of Economics*, 110(3): 605–640, 1995.

[19] Itzhak Gilboa and David Schmeidler. *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, 2001.

[20] Itzhak Gilboa and David Schmeidler. Inductive inference: An axiomatic approach. *Econometrica*, 171(1): 1–26, 2003.

[21] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[22] David Hume. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, 1748.

[23] Richard Jeffrey. *Subjective Probability: The Real Thing*. Cambridge, Cambridge University Press, 2004.

[24] Peter Klibanoff, Massimo Marinacci, and Sujoy Mukerji. A smooth model of decision making under ambiguity. *Econometrica*, 73(6): 1849–1892, 2005.

[25] Frank H. Knight. *Risk, Uncertainty, and Profit*. Boston, New York: Houghton Mifflin, 1921.

[26] David M. Kreps. *Notes on the Theory of Choice*. Westview Press, Boulder, Colorado, 1988.

[27] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[28] Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, Cambridge, 1965.

[29] Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6): 1447–1498, 2006.

[30] John McCarthy. Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2): 27–39, 1980.

[31] Drew McDermott and John Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1–2): 41–72, 1980.

[32] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1): 71–87, 1986.

[33] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3): 241–288, 1986.

[34] Frank P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Harcourt, Brace and Company, New York, 1931.

[35] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2): 81–132, 1980.

[36] Christopher K. Riesbeck and Roger C. Schank. *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1989.

[37] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, New York, 1972 (originally 1954).

[38] Roger C. Schank. *Explanation Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1986.

[39] David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3): 571–587, 1989.

[40] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[41] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London and New York, 1986.

[42] Ray J. Solomonoff. A formal theory of inductive inference I,II. *Information Control*, 7(1,2): 1–22, 224–254, 1964.

[43] Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1922.