Data, Cases and States[*]

Jürgen Eichberger Ani Guerdjikova[†‡]

February 24, 2024

Abstract

In this paper, we provide a novel framework for decision making under uncertainty based on information available in the form of a data set of cases. A case contains information about an action taken, an outcome obtained, and other circumstances that were recorded with the action and the outcome. The set of actions, the set of outcomes and the set of possibly relevant recorded characteristics are derived from the cases in the data set. The information from the data set induces a belief function over outcomes for each action. From a decision maker's preferences over these data-generated belief functions one can derive a representation evaluating outcomes according to the α-maxmin criterion. New data affect behavioral parameters, such as awareness, ambiguity and ambiguity attitude, and may suggest a classifications of data into states. Applications to machine learning and in particular, classification problems are discussed.

Keywords partial information, case-based decisions, data, objective ambiguity, subjective ambiguity attitudes

JEL Classification: D81

[*] Preliminary and incomplete. We would like to thank Itzhak Gilboa and Stefania Minardi for helpful discussions in the early stages of this paper. We also benefited from comments by Edi Karni and Stefan Gerdjikov. We gratefully acknowledge the numerous helpful comments on a previous version of this paper from audiences at RUD 2021 (online), a Research Seminar 2022 (UGA, Grenoble), Theoretischen Ausschuss 2022 (Berlin), D-TEA 2022 (Paris), FUR 2022 (Ghent), Annual Meeting of Verein für Socialpolitik 2022 (Basel), and, in particular, at the Conference in honor of Adi Karni 2022 (Johns Hopkins). This research was funded by the French National Research Agency and the Deutsche Forschungsgemeinschaft (DFG, GermanResearch Foundation – Projektnummer: 529035818) in the French-German research project DATA-AWARE.

[†] Alfred Weber Institute, Department of Economics, University of Heidelberg, Bergheimer Straße 58, 69115 Heidelberg, Germany (email: juergen.eichberger@awi.uni-heidelberg.de) and School of Economics and Finance, HSE (Perm)

[‡] Department of Economics, University of Grenoble-Alpes, GAEL, CNRS, INRA, IUF, 1241 rue des Résidences, 38400 Saint Martin d'Hères, France, email: ani.guerdjikova@univ-grenoble-alpes.fr.

1 Introduction

In economic theory, uncertainty about outcomes of actions is almost always modeled by a set of "states" that together with the action chosen determines the outcomes that the decision maker might experience. Since it is not known which state will be realized when the action is chosen, the decision maker faces uncertainty over the outcome of her actions.

If a probability distribution over the unknown states is known, then each action induces a probability distribution over the outcomes associated with an action. If a few intuitive axioms govern a decision maker's preferences over actions then, as von Neumann and Morgenstern (1944) show, the preference order can be represented by the expected value of the utility of the outcomes obtained from the action.

If probabilities of the states are unknown, Savage (1954) could show that a somewhat more elaborate set of axioms will reveal a "subjective" probability distribution over states that will allow to represent the decision maker's preferences over actions by the expected utility with respect to this subjective probability distribution. Over the past thirty years, research following Schmeidler (1989) and Gilboa and Schmeidler (1989) has derived alternative representations of preferences over actions under weaker axioms that leave room for ambiguity about the true probability distribution over states.

Regarding knowledge about "states", however, most of the literature on decision making under uncertainty assumes that they are an exogenously known feature of the economic model. This view was never uncontroversial (see, e.g., Savage, 1954). In particular, it is challenged by the recent literature on unforeseen contingencies. While learning the probability of states was assumed to occur from data, states where always treated as ex-ante given and unrelated to actually observed data. In this view, data is only collected with respect to well-specified states.

Most statistical data, however, have been collected for purposes completely unrelated to states that uniquely determine the outcomes of actions. Data were collected traditionally for administrative purposes such as military or taxation purposes.[1] Even in financial markets, most data are collected as trading records for the collection of fees or for legal purposes. In general, data are mostly collected for reasons not directly related to the outcomes of actions.

With increasing digital data collection, storing and retrieving facilities, the possibility to relate outcome data to data about actions and related circumstances arises naturally. Methods like data mining and pattern recognition were invented and have been used to extract state-like contingencies from existing data.

Traditionally, economic theory provides little clues for identifying "states" that determine the outcomes of actions. Quite often, states are described only relative

[1] Probably the first census data in England were collected in the "doomsday book" for taxation purposes. Trading data (prices, quantities, offers) in markets were collected for financial contracts, and data about infections were needed to control illnesses. Answers to questionnaires and digital pictures of persons were collected in order to record opinions and movements of people.

to their consequences for given actions such as “good state” or “bad state” according to whether an action produces a “high” or a “low” monetary outcome.

While we do not want to deny that abstract theoretical considerations may help to identify states that determine the outcomes of actions, in this paper, we want to suggest that existing data help to shape and identify circumstances that may determine the outcomes of an action. There may be little relevant data in a new decision situation, yet for more familiar decision situations in the light of more data not only probabilities of outcomes but also the description of circumstances may be revised.

For these reasons, we take a data set of observed cases, rather than a set of states, as the primitive concept of the model of decision making. Information from the data set is the common feature of individual decisions at a given point in time. Growing data may improve information leading not just to updated probabilities over outcomes but eventually to descriptions of states. As we will show in the paper, actions translate to mass distributions over outcomes and preferences over actions to preferences over mass distributions. Hence, the evaluation of actions will be subjective as reflected by the parameter of the representation we derive. Embedding the analysis of data-based action choice in the context of the theory of decision making under uncertainty allows us to apply the theory to economic decision problems.[2]

Decision theory in the tradition of Savage (1954) considers an exogenous set of states of the world and an independent exogenous set of outcomes as primitive concepts of the theory. The state space is exhaustive and observing a state resolves all uncertainty regarding the outcome of an action. Probabilities are derived from a subjective preference over acts. Factual information regarding the set of states and the outcomes of actions as well as the frequencies of states do not enter the description of a decision maker’s choice situation.

Experimental evidence suggests, however, that information about the set of states and the frequency of observations influences choice behavior.[3] Furthermore, the assumption that states are observable is far from innocuous, Gilboa *et al.* (2020). Moreover, while the state space might in principle differ for different decision makers, in economic models, one often reasonably assumes that all agents agree on the state space, possibly because all observe the same data. Such an assumption is however hardly ever made explicit.

To address these issues, we consider, in the spirit of Gilboa and Schmeidler (2001), a data set of cases observed in the past as basis of our theory. Cases in the data set record actions, outcomes and characteristics (circumstances) of decisions observed in the past. Characteristics are factors that influence or determine the set of outcomes of an action. They are observable, can be retrieved from data, and can serve as empirical proxies for states.

Characteristics differ from the Savage (1954) concept of states, in three important aspects. First, a characteristic may fail to specify an outcome for all available

[2] In contrast to Billot *et al.* (2005) and Eichberger and Guerdjikova (2010).

[3] The famous paradoxes of Ellsberg (1961) suggest that partial information about the probability of events substantially influences subjects’ choices.

actions, e.g., because it has never been observed in combination with this action. Second, the set of characteristics observed in the data need not be exhaustive: it may capture only a subset of all relevant contingencies. Third, characteristics may provide a coarse description of the underlying uncertainty with a single characteristic corresponding to a set of states. Taking characteristics as a primitive for our approach means that the relevant factors for determining the outcomes of an action need not and cannot be specified ex-ante before the analysis can begin.

Since data do not uniquely identify the relevant state space[4], subjective factors will influence the decision maker's evaluation of actions: (i) predictions about counterfactuals; (ii) awareness of "other, yet unobserved" but relevant characteristics [5], or of yet unidentified but relevant categories, and (iii) attitude to the indeterminacy of predictions given such unawareness.

Formally, the incompleteness in the data implies that each action induces a set of possible outcomes for a given characteristic. Combined with the observed frequencies of characteristics, we obtain a mass distribution (belief function) over sets of outcome distributions. Subjective preferences over such mass distributions induce a representation of preferences over actions as in Jaffray (1989). Notably, the attitude towards uncertainty captured by an optimism parameter determines the evaluation of set-valued outcomes.

In the first part of the paper, we axiomatize a preference representation for a given data set. This representation combines the objective information in the data with the subjective characteristics of the decision maker. It identifies the perception of unawareness: a subjective probability γ for the possibility that an "other" so far unobserved characteristic may occur. We also identify the subjective attitude towards unawareness: captured by a coefficient of optimism. We also address the issue of subjective predictions about counterfactuals. We propose a method to determine those cases in the data which are considered most relevant for the evaluation of the choice of an action for a given characteristic. We then derive the sets of subjective predictions as the set of observed frequencies for the most relevant cases.

In the second part of the paper, we study the responses as new data become available. When new data simply increase the number of observations of cases already considered, statistical learning about frequencies arises naturally as a special but well-defined case of our approach. In contrast to the state-based approach, the distinction between objective, i.e., data-based, information and subjective perception of uncertainty provides a framework for studying learning about the relevant state-space. The occurrence of action-characteristic pairs unobserved before replaces the uncertainty about counterfactuals with objective outcome distributions similar to theories introduced in Karni (2022). Observation of new characteristics (similar in spirit though not in detail, to the approaches by Karni and Vierø (2017) and by Gilboa *et al.* (2020)) leads to an expansion of the perceived state space.

[4] Few decision situations specify precisely the mapping from actions to state-contingent outcomes. Bets as the prototypes of Savagean acts are discussed in Section 2.2.1.

[5] This is similar to unimaginable consequences or actions of which the decision maker is unaware of but that determine new conceivable states in Karni and Vierø (2017, p. 304).

The discovery of new categories can provide a refinement for an initially coarse perception of the state space. Furthermore, evidence from data can be used to update the decision maker's subjective perception of uncertainty (unawareness) as well as his attitude towards it. Our framework thus provides a way of modeling dynamic awareness of the state space which evolves with the available data and might approach the ideal of a Savagean state space.

Our framework can be used to provide decision-theoretic foundations for models of machine learning in artificial intelligence. The classification problem described in Section 2.2.2 illustrates how our proposed representation can be used to evaluate algorithms. While commonly used methods, such as entropy minimization can be obtained as special cases of the representation, our approach specifically takes into account the possible incompleteness of data and identifies the subjective factors (parameters) necessary for the evaluation of algorithms. Finally, our setting allows us to take into account complexity as reflected by the description of the set of characteristics. The identification of relevance classes pinpoints the subjective categorization used by the decision maker on the set of characteristics in the classification problem. While using a finer categorization allows for more precise predictions conditional on a given characteristic, it also increases the number of observations necessary to confidently learn the underlying probabilities. A decision maker who exhibits a coarser categorization can be deemed to have a higher cost of complexity. In a dynamic setting in which new categories become available, the choice of which categories to use and which to ignore is similar to the problem of structural risk minimization in statistical learning theory. Our framework provides a way to explore these issues from a decision-theoretic point of view.

2 Concepts, notation and leading examples

In this paper, we will not assume a priori known sets of actions and consequences. In contrast to most of the literature, we will derive these sets from a data set of previously observed cases. We first introduce the main concepts and notation.

2.1 Cases and states: the basic model

The primitive concept of our approach is a case $c = (a, x, r)$ that records an *action* a, an *outcome* r, and a vector of *characteristics* x listing possibly relevant context variables. *Information* available at the point of decision making is a finite data set of cases that have been observed and recorded in previous decision situations:

$$D = \left\{ (a_n, x_n, r_n)_{n=1}^{N} \right\}.$$

Note that the same case $c = (a, x, r)$ may have been observed several times. In this paper, we assume that records of cases are complete (no missing entries).

Given a data set of cases D, the set of observed actions is given by:

$$A_D = \{a \mid (a_n = a; x_n; r_n) \in D \text{ for some } n \in \{1...N\}\}.$$

The set of observed outcomes is:

$$R_D = \{r \mid (a_n; x_n; r_n = r) \in D \text{ for some } n \in \{1...N\}\}.$$

Characteristics recording the circumstances of a decision may be classified in categories. For example, a medical doctor who has recorded the case of a patient with a particular treatment usually also notes some biometric characteristics of the patient. We will refer to the type of biometric data recorded, such as blood pressure, temperature, weight, etc, as categories and to the entries in these categories as characteristics. Hence, *categories* classify characteristics. The data set D identifies a set of categories T and, for each category $t \in T$, the set of observed characteristics:

$$X_D^t = \left\{x^t \mid \left(a_n; x_n = \left(x_n^1, .., x_n^t = x^t, ..x_n^T\right); r_n\right) \in D \text{ for some } n \in \{1...N\}\right\}.$$

The set of all characteristics is obtained as the Cartesian product of X_D^t:

$$X_D = \prod_{t=1}^{T} X_D^t.$$

When we refer to characteristics without mentioning a category, we mean the vector $x \in X_D$ with components x^t for all categories $t \in T$. The recorded characteristics may simply reflect the nature of the available data or be deliberately chosen to reflect a theory about the factors influencing the action payoffs.

The data set D also specifies for each characteristic $x \in X_D$ the frequency with which this characteristic has been observed

$$f_D(x) = \frac{|\{(a_n, x_n, r_n) \in D \mid (a_n, x_n, r_n) = (a_n, x, r_n)\}|}{N}.$$

Typically, in a data set D, the outcome observed from an action $a \in A_D$ in combination with characteristic $x \in X_D$ will not be unique. Thus, we associate with a pair (a, x) the conditional frequency $\rho_D(r \mid a, x)$ of an outcome $r \in R_D$ when action $a \in A_D$ is chosen and characteristic $x \in X_D$ has been realized:

$$\rho_D(r \mid a, x) = \frac{|\{(a_n, x_n, r_n) \in D \mid (a_n, x_n, r_n) = (a, x, r)\}|}{|\{(a_n, x_n, r_n) \in D \mid (a_n, x_n) = (a, x)\}|}.$$

Denoting by $\Delta(Z)$ the simplex of all probability distributions over a finite set Z, we have $f_D \in \Delta(X_D)$ and $\rho_D(\cdot \mid a, x) \in \Delta(R_D)$. We will denote by $\Re_D$ the set of all finite subsets of $\Delta(R_D)$.

Remark 1. We emphasize that, for a given point in time, the decision maker's information is fully summarized by the data set D, which we take as given. In particular, there is no prior information about the set of possible categories, their relevance, or the number of characteristics within each category. Over time with new data, however, new actions, new outcomes, new categories, or new characteristics may be discovered, as discussed in Section 5.

2.1.1 Ambiguity: Uncertainty about outcome distributions

While a state identifies the outcome of each available action, this need not be the case for a characteristic. If an action-characteristic combination (a, x) has not been observed[6] in D, the corresponding frequency of outcomes $\rho_D(\cdot \mid a, x)$ is not well defined. The decision maker will thus have to make a subjective prediction about the outcome of a when characteristic x occurs.

Different methods can be used to arrive at such predictions: statistical methods, logical inference, analogy or similarity. In certain situations, such a method might uniquely identify the distribution of outcomes. In general, however, a set of possible distributions will obtain,[7] see the examples in Section 2.2. The following sets of outcome distributions appear to be natural candidates for unobserved action-characteristics pairs (a, x): (i) the set of all frequencies over outcomes in the data D, $\mathbf{R}_D = \cup_{(a,x)\in A_D\times X_D}\mathbf{R}_D(a, x)$; (ii) the set of all frequencies over outcomes observed in combination with a particular action a, $\mathbf{R}_D(a) = \cup_{x\in X_D}\mathbf{R}_D(a, x)$; (iii) the set of all frequencies over outcomes observed in combination with a particular characteristic x, $\mathbf{R}_D(x) = \cup_{a\in A_D}\mathbf{R}_D(a, x)$. Note that the sets of outcome distributions $\mathbf{R}_D, \mathbf{R}_D(a), \mathbf{R}_D(x)$ are all finite subsets of $\Delta(R_D)$. More generally, the decision maker may, for each data set D and each (a, x) identify a subset of the observed action-characteristic pairs – the most relevant observations for (a, x) – and use the corresponding observed frequencies of outcomes to form the set $\mathbf{R}_D(a, x)$. Section 4 describes how the most relevant sets can be identified.

We denote by $\mathbf{R}_D(a, x) \subset \Delta(R_D)$ the set of possible outcome distributions the decision maker associates with (a, x), and assume that this set is finite, i.e. $\mathbf{R}_D(a, x) \in \Re_D$. These sets are subjective but data-based. For observed (a, x)-combinations, this set is a singleton and comprises the observed frequency of outcomes $\mathbf{R}_D(a, x) = \{\rho_D(\cdot \mid a, x)\}$. We will maintain the following assumptions for all $(a, x) \in A_D \times X_D$:

- experience-based beliefs: $\mathbf{R}(a, x) \subseteq \mathbf{R}_D = \cup_{(a,x)\in A_D\times X_D}\mathbf{R}_D(a, x)$;
- data-based beliefs: $\rho_D(\cdot \mid a, x) \in \mathbf{R}(a, x)$;
- possibility of degenerate distributions: if $\rho \in \mathbf{R}(a, x) \implies 1_{\{r\}} \in \mathbf{R}(a, x)$ for all $r \in \operatorname{supp}\rho$.

For most of the discussion, we assume that the decision maker does not perceive ambiguity due to the limited number of observations of a specific (a, x)-combination. Such ambiguity is discussed in Eichberger and Guerdjikova (2010) and Eichberger and Guerdjikova (2013), in the context of Billot *et al.* (2005). However, using the methods applied in Eichberger and Guerdjikova (2013), it is easy to incorporate the dependence of subjective predictions on the amount of relevant data available. We explain how this can be done in Remark 4 in Section 4.

[6] Missing counterfactuals may occur even in large data sets D because of practical, legal or moral constraints on actions.

[7] E.g., a non-parametric model might be only partially identified; the decision maker may decide to use the confidence interval of a parametric estimation instead of the estimate itself; there might be uncertainty about the correct analogy, etc.

2.1.2 Awareness of unawareness: Other characteristics

While the set of Savage states is exhaustive, i.e., describes all relevant contingencies, the set of observed characteristics need not be: there might be a state which does not correspond to any of the characteristics observed in the data, X_D. It may be that such characteristics cannot occur in the context in which the data were collected or that there have not yet been sufficient observations. Recording categories of decision-relevant characteristics may make a decision maker aware of the fact that some category could contain yet-unobserved characteristics.

We model the awareness of "other, so far unobserved, characteristics" by extending the set of characteristics X_D with a (place holder) characteristic "x_o" (for "other characteristics"). The augmented set of characteristics is $\hat{X}_D = X_D \cup \{x_o\}$.

From the data set D, neither a frequency for such "other" characteristics nor an outcome distributions $\rho_D(\cdot \mid a, x_o)$ can be deduced. A decision maker thus faces ambiguity given this lack of information and may again associate a set of distributions $\mathbf{R}_D(a, x_o)$ with the occurrence of x_o. A possible candidate is the set of all outcome distributions that have been observed for an action in D, $\mathbf{R}_D(a, x_o) = \cup_{x \in X_D} \mathbf{R}_D(a, x)$.[8]

The decision maker is assumed to attribute a subjective weight, interpreted as his degree of unawareness, γ_D to this unobserved characteristic. $(1-\gamma_D)$ is then the degree of confidence assigned to the information in the data set and in particular to the frequency of observed characteristics $f_D(x)$. This degree of unawareness is purely subjective and will be derived from the decision maker's preferences.[9]

2.1.3 Awareness of unawareness: Other categories

Each state uniquely identifies the outcome of an action. In contrast, as we saw above, an action a in combination with a given characteristic x can result in several distinct outcomes. In some situation, such variation can be considered as noise and the decision maker might reasonably use the observed frequency of outcomes $\rho_D(\cdot \mid a, x)$ as a unique prediction. In other contexts, the decision maker might infer that the observed variability is due to some underlying, but so far unobserved (latent) factor. He would thus be aware that the characteristic x corresponds to a set of states rather than to a single state and hypothesize the existence of a yet unobserved category, the characteristics of which would correspond to the underlying states grouped in x. This can result in ambiguity in the immediate

[8] Below, we allow also for general sets of outcome distributions. Chung *et al.* (2018) provide a method of estimating the outcomes on "unknown unknowns" in the context of machine learning. Karni (2022) models "theories" about possible outcome distributions.

[9] Although unknown in the data set at a particular point in time, a sequence of data sets may reveal information about the frequency of unobserved characteristics. The degree of unawareness can, e.g., be related to the frequency of new characteristics observed over time. We consider such learning in Section 5. Chung *et al.* (2018) provide an econometric method for estimating the weight assigned to "unknown unknowns" in the context of classification, while Schipper (2022) suggests a behavioral approach in a EU-setting.

sense of the word.[10]

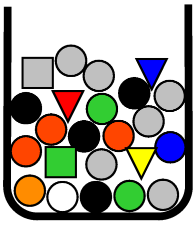
2.2 Leading examples

Before proceeding to the analysis of choice under uncertainty in this framework, we will illustrate our suggested approach by two examples: the classic situation of betting on an urn, and a medical decision.

2.2.1 Urn with unknown content

Consider an urn containing unknown objects. Sequentially, objects are drawn from the urn. For each object a list of properties (characteristics) is recorded in the data D. Such characteristics could include

- **color**: red, blue, yellow, ...
- **shape**: ball, cube, pyramid,.....
- **material**: wood, iron, glass, ...
- **weight** (grams): 20,10,50, ...



Characteristics are classified into categories: color c, shape s, material m, and weight w. A category is a set of characteristics of the same type, such as a set of colors or a set of shapes. Hence, a characteristic registered in a case may be a quadruple (x^c, x^s, x^m, x^w) indicating the color, the shape, the material, and the weight of the object drawn from the urn in this case.

Actions are bets on characteristics of the next object drawn from the urn. Outcomes are monetary payments $r \in \mathbb{R}$. All information of the agent is given by a data set of N observed cases:

$$D = \left\{ (a_n, x_n, r_n)_{n=1}^{N} \right\}.$$

Example 1. Table 1 shows a data set $D = \left\{ (c_n)_{n=1}^{20} \right\}$ of 20 cases with four actions, $A_D = \{a_1, a_2, a_3, a_4\}$, two outcomes, $R_D = \{0, 1\}$ and a single category, the color of the objects, $X_D = \{R, B, Y\}$. Table 2 organizes these cases in a matrix listing the observed outcome distributions with respect to the actions and the characteristics. The first table records the first 10 cases and the second table all 20 cases. A tuple (z_1, z_2) records the frequency of outcome $r = 0$ by z_1 and the frequency of outcome $r = 1$ by z_2. If an action-characteristic pair (a, x) has not been observed yet, there is ambiguity about the outcome distribution with a set of possible distributions $\mathbf{R}_D(a, x)$. The more cases are contained in the data set the fewer cells of the matrix will be left open. We also include a column for so far unobserved colors, characteristics x_o the decision maker might be unaware of.

For a bet where the winning condition is recorded as a characteristic, e.g. a_2 means "betting on B" and B is a characteristic in X_D, it appears natural to

[10] Cicero writes "ex ambiguous controversial nascitur, cum res in unam sententiam scripta duas aut plures sententias significat".(Short, 2018, p.3)

D	A	X^1	R
c_1	a_2	Y	1
c_2	a_1	B	0
c_3	a_3	Y	1
c_4	a_4	Y	1
c_5	a_2	B	1
c_6	a_3	R	1
c_7	a_4	B	1
c_8	a_4	B	1
c_9	a_1	Y	0
c_{10}	a_2	Y	0

D	A	X^1	R
c_{11}	a_2	B	1
c_{12}	a_1	Y	0
c_{13}	a_3	B	0
c_{14}	a_2	B	1
c_{15}	a_1	R	1
c_{16}	a_3	Y	1
c_{17}	a_4	R	0
c_{18}	a_1	R	1
c_{19}	a_3	B	0
c_{20}	a_2	Y	0

Table 1: Data set: $N = 20$

	$(1-\gamma)f_D(R) = \frac{(1-\gamma)}{10}$	$(1-\gamma)f_D(B) = \frac{4(1-\gamma)}{10}$	$(1-\gamma)f_D(Y) = \frac{5(1-\gamma)}{10}$	γ
a_1	$\mathbf{R}_D(a_1, R)$	$\{(1,0)\}$	$\{(1,0)\}$	$\mathbf{R}_D\,(a_1, x_o)$
a_2	$\mathbf{R}_D(a_2, R)$	$\{(0,1)\}$	$\{(\frac{1}{2}, \frac{1}{2})\}$	$\mathbf{R}_D\,(a_2, x_o)$
a_3	$\{(0,1)\}$	$\mathbf{R}_D(a_3, B)$	$\{(0,1)\}$	$\mathbf{R}_D\,(a_3, x_o)$
a_4	$\mathbf{R}_D\,(a_4, R)$	$\{(0,1)\}$	$\{(0,1)\}$	$\mathbf{R}_D\,(a_4, x_o)$

$N = 10$

	$(1-\gamma)f_D(R) = \frac{4(1-\gamma)}{20}$	$(1-\gamma)f_D(B) = \frac{8(1-\gamma)}{20}$	$(1-\gamma)f_D(Y) = \frac{8(1-\gamma)}{20}$	γ
a_1	$\{(0,1)\}$	$\{(1,0)\}$	$\{(1,0)\}$	$\mathbf{R}_D\,(a_1, x_o)$
a_2	$\mathbf{R}_D\,(a_2, R)$	$\{(0,1)\}$	$\{(\frac{2}{3}, \frac{1}{3})\}$	$\mathbf{R}_D\,(a_2, x_o)$
a_3	$\{(0,1)\}$	$\{(1,0)\}$	$\{(0,1)\}$	$\mathbf{R}_D\,(a_3, x_o)$
a_4	$\{(1,0)\}$	$\{(0,1)\}$	$\{(0,1)\}$	$\mathbf{R}_D\,(a_4, x_o)$

$N = 20$

Table 2: Data sets: $N = 10$ and $N = 20$

assume that outcome distributions are concentrated, i.e., $\rho_D(a_{2,},B) = (0,1)$ and $\rho_D(a_{2,},x) = (1,0)$ for $x \neq B$. If the action is, however, not specified completely in regard to the characteristics in the data, for example, if action a_2 is a "bet on the color Y and the shape cube, while the set of characteristics X_D records only colors, then outcome frequencies need not be concentrated on $r = 0$ or $r = 1$. Instead, non-degenerate outcome frequencies $\left(\frac{2}{3}, \frac{1}{3}\right)$ will occur, reflecting the missing category of "shapes". Outcome distributions that are not Dirac measures might make the decision maker aware of missing categories.

Example 1 shows the distinction between our framework and typical experiments in statistics, which rely on the specification of states. The latter consider an urn for which it is known that all objects are balls that are distinct only in color. Furthermore, the set of possible colors is specified. It is thus known that there is a single category "color" with an exhaustive list of possible characteristics. The only unknown aspect concerns the frequency of the colors in the urn. The Savagean acts are bets on colors which are known to be in the urn. Learning the color of the ball drawn from the urn resolves all uncertainty and determines uniquely the outcome for all acts. Each characteristic, i.e., each "color", is a state in the sense of (Savage, 1954).

2.2.2 An Application to Machine Learning: The Classification Problem

Consider the problem of choosing a procedure **a**, an *algorithm*, for classifying a finite set of objects Ω into a finite set of classes K based on data consisting of a vector of characteristics x of the objects. Without loss of generality, assume that there is an (unknown) function $\kappa : \Omega \to K$ that indicates for each object $\omega \in \Omega$ the class $\kappa(\omega) \in K$ it belongs to. In order to predict the class to which an object ω belongs, the data contain a vector of characteristics $\chi(\omega) \in X$ associated with the object ω. An **algorithm a** is a (computer) program that predicts for an object ω with characteristics $x \in X$ (the likelihood of) its class $k \in K$ based on the data in the set D. An algorithm, specifies the choice of an action for each data set: $\mathbf{a} : \mathbb{D} \to \Delta(K)$. If $\mathbf{a}(D) = a$, we write $a(k, x)$ for the probability assigned by the algorithm to class k given characteristics x and data set D. Algorithms thus operate on data directly and generate outcomes. Therefore, we will take the algorithm not as part of a case in the data but consider for the different algorithms sets of data that contain data points (examples) classified by their labels only. Each algorithm thus identifies a procedure for classifying the objects according to their characteristics (labels) $x \in X$.

The algorithm derives all information about the features (characteristics of an object) from a finite **data set** $\hat{D}$. The data set $\hat{D}$ contains *data points (examples)* of objects for which the characteristics x have been recorded. For most data points only these features x have been recorded. One distinguishes

- *training data* as in the data set $D \subset \hat{D}$ that is correctly labeled and can be used to determine the parameters of the algorithm and

- *test data* that contains objects and their features (characteristics) but no labels. Denote by $\tilde{D} = \{x_i, i = 1, ..., M\} := \hat{D} \backslash D_e$ the set of M unlabeled examples for which only the features x_i are given in the data set $\tilde{D}$.

As before, we assume that the set of characteristics $X_{\hat{D}}$ is derived from the data set $\hat{D}$. As above, characteristics are taken to be vectors with each entry corresponding to a category (feature of the objects), $x_i = (x_i^t)_{t \in T_{\hat{D}}}$. Note that not all characteristics in $X_{\hat{D}}$ need occur in the labeled training set, $X_D \subseteq X_{\hat{D}}$.

The data set $\hat{D}$ provides objective information about the frequency of characteristics in $X_{\hat{D}}$, $f_{\hat{D}}(x)$. The training data set D identifies the frequency of an object with characteristics x being classified as k, $f_D(x, k)$.

We thus obtain from D a classification likelihood $l_D(x) : X_D \to \Delta(K)$, the relative frequency with which an object with vector of features x is classified as belonging to class k:

$$l_D(x,k) = \left(\frac{f_D(x,k)}{\sum_{\tilde{k}\in K} f_D\left(x,\tilde{k}\right)} \right)_{k\in K} = \left(\frac{|(x_i,k_i) \in D \mid x_i = x, k_i = k|}{|(x_i,k_i) \in D \mid x_i = x|} \right)_{k\in K}$$

The choice of algorithm is related to the task under consideration and is partly a matter of subjective assessment.

> "Nearly all deep learning algorithms can be described as particular instances of a fairly simple recipe: combine a specification of a data set, a cost function, an optimization procedure and a model." Goodfellow *et al.* (2016, p.151)

[11]

The payoff of an algorithm for a given object with characteristics x is given by the probability with which it correctly classifies the object, $r \in [0,1]$[12]. For any algorithm $\mathbf{a}$, one can assign to any **labeled data set** D such that $\mathbf{a}(D,x) = a(k,x)$, the payoff distribution $\rho_D(a(k,x),x) = \rho_D(\mathbf{a}(D,x),x)$:

$$\rho_D(r \mid a(k,x),x) = \sum_{k\in K|a(k,x)=r} l_D(x,k)$$

Applying an algorithm $\mathbf{a}$ to an example x_i from the test data $\tilde{D}$ such that $\mathbf{a}(D) = a(k,x)$ yields a likelihood over predicted classes but no outcome, since the objects in $\tilde{D}$ are not labeled.

In order to assess the predictive quality of an algorithm $\mathbf{a}$ trained by the labeled examples in the training set D when applied to examples of the test data, one needs to make assumptions about the classification $k \in K$ of the objects in the set $\tilde{D}$ conditional on its features x. Assuming that the elements of $\tilde{D}$ are independently drawn from a data generating process with probability distribution $p(k,x)$, one obtains for $\mathbf{a}(D,x) = a(k,x)$

$$\rho_{\tilde{D}}(r \mid a(k,x),x) = \sum_{k\in K|a(k,x)=r} p(k \mid x).$$

One usually assumes that labeled examples in the training data D were also (independently) drawn from the same data generating probability distribution $p(k,x)$. Unless the number of observations in both D and $\tilde{D}$ is large, there is, however, no reason to assume that the probability of a correct prediction for an object with features x, $\rho_D(x,a)$ is the same for examples from the training set D as for examples with the same features x from the test set $\tilde{D}$, $\rho_{\tilde{D}}(x,a(k,x))$. Since

[11] Algorithms can be obtained from regression, density estimation, probability mass function estimation, and many other procedures etc. Goodfellow *et al.* (2016, pp. 98-101, 137-161).

[12] This assumes that the decision maker does not consider some classifications as more important than others. Here, as in most classification algorithms, we distinguish only correcly classified objects from falsely classified cases

the labeled cases in the training set are used to determine the parameters of the algorithm $\mathbf{a}$, in general, a given algorithm $\mathbf{a}$ will make correct classifications more often for examples from the training set than for examples from the test data.

This fact creates a well-known conflict between over- and under-fitting in the evaluation of an algorithm. There is no optimal solution to this conflict Goodfellow *et al.* (2016, Section 5.2, pp. 108-118) but one can take this conflict into account when evaluating the outcome of an algorithm.[13]

Our approach addresses this conflict explicitly by considering sets of outcome distributions: $\mathbf{R}_D\left(a,x\right)=\left\{\rho_D\left(x,a\left(k,x\right)\right),\rho_{\tilde{D}}\left(x,a\left(k,x\right)\right)\right\}$ for a given pair of algorithm $\mathbf{a}$, with $\mathbf{a}\left(D,x\right)=a\left(k,x\right)$, and features x.

A second difficulty concerns the classifications of characteristics which occur in the set of unlabeled examples $\tilde{D}$, but not in the training set D, $X_{\hat{D}}\backslash X_D$. For such characteristics, the set $\mathbf{R}_D\left(a\left(k,x\right),x\right)$ has to be specified without recurring to available data about x. Common procedures in machine learning include kernel and nearest neighbor methods and consist in using the observed frequencies of "similar" characteristics, x'. Defining similarity or relevance, however, is a subjective judgment which has to be made based on the context and prior knowledge about the problem. In Section 4, we show how such subjective perception of relevance can be identified based on preferences over algorithms. This in turn identifies subjective assignments of sets of likelihoods to characteristics observed only in unlabeled examples, $\mathbf{L}_D\left(x\right)$, and the corresponding sets of outcome distributions, $\mathbf{R}_D\left(a\left(k,x\right),x\right)$.

Given the frequency of characteristics $f_{\hat{D}}\left(x\right)$ observed in the data, an algorithm $\mathbf{a}$ can be identified with a mass distribution $m_{\mathbf{a}}^{f_{\hat{D}}}$, which assigns to each set of outcome distributions R the frequency of those characteristics for which R occurs.

$$m_{\mathbf{a}}^{f_D}\left(R\right)=\sum_{\left\{x\in X_D|\mathbf{R}_D\left(a\left(k,x\right),x\right)=R\right\}}f_D\left(x\right)$$

Thus, an algorithm $\mathbf{a}$ can be viewed as a probability distribution over the finite sets of outcome distributions.

Below, we propose axioms inspired by those advanced in Jaffray (1989) yielding an α-maxmin evaluation of algorithms. Suppose, in particular, that the utility function over outcomes is logarithmic, $u\left(r\right)=\ln r$. For a given data set $\hat{D}$ and a corresponding training data-set D and test data-set, the algorithm $\mathbf{a}$ such that $\mathbf{a}\left(D,x\right)=a\left(k,x\right)$ is evaluated as:

$$V_{\hat{D}}\left(a\right)=\sum_{x\in X_{\hat{D}}}f_{\tilde{D}}\left(x\right)\left[\alpha_D\max_{\rho\in\mathbf{R}_D(a,x)}\sum_r\rho\left(r\right)\ln r+\left(1-\alpha_D\right)\min_{\rho\in\mathbf{R}_D(a,x)}\sum_r\rho\left(r\right)\ln r\right]$$

or, using the observed likelihoods,

[13] Including, weight decay in addition to the mean squared error for comparing the bias of the estimator with the variance of the parameter estimates Goodfellow *et al.* (2016, p.127).

$$V_{\hat{D}}\left(a\right)=\sum_{x\in X_{\hat{D}}}f_{\tilde{D}}\left(x\right)\left[\begin{array}{c}\alpha_D\max_{l_D(x)\in\mathbf{L}_D(x)}\sum_{k\in K}l_D\left(k,x\right)\ln\left(a\left(k,x\right)\right)\\+\left(1-\alpha_D\right)\min_{l_D(x)\in\mathbf{L}_D(x)}\sum_{k\in K}l_D\left(k,x\right)\ln\left(a\left(k,x\right)\right)\end{array}\right],$$

where

- $\alpha_D\in[0,1]$ is the parameter of optimism, $(1-\alpha_D)$ is the degree of pessimism given the observed training data D;
- $\gamma_D\in(0,1)$ is the subjective degree of ambiguity.

In the special case in which the number of observations is large and the sets of characteristics in the training and in the test data-set coincide, $X_D=X_{\hat{D}}$, the sets of likelihoods can be taken to be singletons and coincide with the observed frequencies, $\mathbf{L}_D\left(k,x\right)=l_D\left(k,x\right)$. For this special case, the evaluation of the algorithm is based on its relative entropy:

$$V_{\hat{D}}\left(a\right)=\sum_{x\in X_D}f_{\tilde{D}}\left(x\right)\sum_k l_D\left(k,x\right)\ln\left(a\left(k,x\right)\right). \tag{1}$$

Note that $-V_D\left(a\right)$ is the entropy of the algorithm relative to the observed frequency of characteristics and the classification likelihoods. The entropy is minimized by setting $a\left(k,x\right)=l_D\left(k,x\right)$ for all k and x.

Finally, the possibility of other (potentially unobserved characteristics) has to be taken into account, when the algorithm is applied beyond the test data set $\tilde{D}$. For such characteristics, x^o, the sets $\mathbf{L}_D\left(x^o\right)$ and $\mathbf{R}_D\left(\mathbf{a},x^o\right)$ also have to be subjectively specified, and a weight γ_D has to be assigned to the observation of such "other" characteristics. The resulting representation is:

$$\begin{aligned}V_{\hat{D}}\left(a\right)&=&\left(1-\gamma_D\right)\sum_{x\in X_{\hat{D}}}f_{\tilde{D}}\left(x\right)\left[\alpha_D\max_{\rho\in\mathbf{R}_D(a,x)}\sum_r\rho\left(r\right)\ln r+\left(1-\alpha_D\right)\min_{\rho\in\mathbf{R}_D(a,x)}\sum_r\rho\left(r\right)\ln r\right]\\&&+\gamma_D\left[\alpha_D^o\max_{\rho\in\mathbf{R}_D(a,x^o)}\sum_r\rho\left(r\right)\ln r+\left(1-\alpha_D^o\right)\min_{\rho\in\mathbf{R}_D(a,x^o)}\sum_r\rho\left(r\right)\ln r\right]\end{aligned}$$

where α_D^o is the degree of optimism for unobserved characteristics.

3 Decisions for a given set of data

In this section, we show that the data-based framework that we introduced generates a belief function over outcome distributions for each action. Hence, one can derive a representation of preferences over these belief functions similar to Jaffray (1989). In addition, we provide axioms in order to characterize a subjective degree of unawareness regarding potential other characteristics.

3.1 From data to choice

At a given point in time, a decision maker knows the data in a set D. As described in Section 2, for each pair of actions and characteristics $(a, x) \in A_D \times X_D$, there is a finite set of outcome distributions $\mathbf{R}_D(a, x) \subset \Delta(R_D)$. Recall that $\mathbf{R}_D := \cup_{a \in A_D, x \in X_D} \mathbf{R}_D(a, x)$ is the finite set of all frequency distributions over outcomes in D. We assume that $\mathbf{R}_D$ contains all degenerate outcome distributions, δ_r for $r \in R_D$. We denote by $\boldsymbol{\mathcal{R}}_D$ the set of all subsets of $\mathbf{R}_D$ and by $\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$the set of all probability distributions on $\boldsymbol{\mathcal{R}}_D$.

The following table summarizes the primitive concepts derived from data in D and the relevant notation:

Summary of basic notation:
A finite data set of cases: $c = (a, x, r) \in D$, induces

- a set of actions: $a \in A_D$,
- a set of characteristics: $x \in X_D$,
  - a frequency distribution over characteristics: $f_D \in \Delta(X_D)$,
- extended set of characteristics: $\hat{X}_D = X_D \cup \{x_o\}$
  - degree of unawareness: γ_D
- a set of outcomes: R_D,
  - a frequency distribution over outcomes: $\rho \in \Delta(R_D)$,
  - the set of all finite subsets of $\Delta(R_D)$: $\Re_D$
  - for each $(a, x) \in A_D \times X_D$, a finite set of frequencies over outcomes: $\mathbf{R}_D(a, x) \in \Re_D$
- the set of all frequency distributions over outcomes in D: $\mathbf{R}_D := \cup_{a \in A_D, x \in X_D} \mathbf{R}_D(a, x)$
  - $\delta_r \in \mathbf{R}_D$ for all $r \in R_D$
  - the set of all subsets of $\mathbf{R}_D$: $\boldsymbol{\mathcal{R}}_D$
  - the set of all probability distributions on $\boldsymbol{\mathcal{R}}_D$: $\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$

3.2 From actions to mass distributions

We first specify actions on the set of observed characteristics, X_D.

For a given action $a \in A_D$, consider the mapping: $a : X_D \rightarrow \boldsymbol{\mathcal{R}}_D$ which to every characteristic x assigns the predicted set of outcomes of action a for this characteristic, $a\left(x\right) = \mathbf{R}_D(a, x) \in \boldsymbol{\mathcal{R}}_D$. Note that the frequency of characteristics in the data $f_D\left(x\right)$ gives a probability distribution over these predictions, assigning a probability of $f_D\left(x\right)$ to $\mathbf{R}_D(a, x)$. Given action a, the observed frequency of characteristics generates a probability distribution over the power set of $\mathbf{R}_D$:

$$m_a^{f_D}(R) = \sum_{\{x \in X_D | \mathbf{R}_D(a,x)=R\}} f_D(x).$$

By definition, $m_a^{f_D}(R) \geq 0$ for all $R \in \boldsymbol{\mathcal{R}}_D$ and $\sum_{R \in \boldsymbol{\mathcal{R}}_D} m_a^{f_D}(R) = 1$.

A probability distribution over the elements of a power set, $m_a^{f_D} \in \Delta\left(\boldsymbol{\mathcal{R}}_D\right)$ is a mass distribution that defines a belief function, Grabisch (2016, p. 380). Since the outcomes of actions on the set of observed characteristics can be represented as a mass distribution, we use the seminal approach by Jaffray (1989) to characterize preferences.

The set of actions A_D together with the realized frequency f_D generates a finite set of mass distributions $\left\{m_a^{f_D} : a \in A_D\right\}$. Similarly to Savage (1954), we assume that the set of hypothetical actions $\mathbf{A}_D$ which the decision maker can conceive is larger than $\left\{m_a^{f_D} : a \in A_D\right\}$ and includes all mappings from characteristics to sets of observed outcome distributions in $\boldsymbol{\mathcal{R}}_D$:

$$\mathbf{A}_D = \left\{a : X_D \to \boldsymbol{\mathcal{R}}_D\right\}.$$

We call an action unambiguous when $a\left(x\right)$ is a singleton for all x. Such actions induce a probability distribution over outcomes $m_a\left(r\right) = \sum_{x \in X_D} f_D\left(x\right) \rho_D\left(r \mid a, x\right)$.

Note that the specification of an action combines the (set-valued) consequences of actions with the information contained in the mass distribution. If this information suffices to associate a probability distribution with each action (when all actions are unambiguous) then preferences will be over probability distributions as in von Neumann and Morgenstern (1944).

The following example illustrates this construction.

Example 2. Consider a data set D with two characteristics $X_D = \{x_1, x_2\}$ yielding two outcome distributions $\mathbf{R}_D = \{\rho_1, \rho_2\}$ with the power set $\boldsymbol{\mathcal{R}}_D = \mathcal{P}(\{\rho_1, \rho_2\}) = \{\{\rho_1\}, \{\rho_2\}, \{\rho_1, \rho_2\}\}$. The set of all basic actions is $\mathbf{A}_D = \{a : X_D \to \boldsymbol{\mathcal{R}}_D\}$. Given a probability (frequency) distribution over the characteristics $\{x_1, x_2\}$, say (f_1, f_2), each action $a \in \mathbf{A}_D$ induces a mass distribution m_a^f in $\Delta(\boldsymbol{\mathcal{R}}_D)$. Given the distribution f on X_D, the nine acts in $\mathbf{A}_D$ induce nine mass distributions $m_a^f \in \Delta\left(\boldsymbol{\mathcal{R}}_D\right)$ as illustrated in Table 3.

As Example 2 illustrates, the set of mass distributions $m_a^{f_D}$ induced by the actions $a \in A_D$ together with the frequency distribution f_D observed in a data set D will be a small subset of all mass distributions $\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$. Allowing for mixtures of acts in $\mathbf{A}_D$, however, will extend the set of mass distributions on $\mathbf{R}_D$ considerably.

For $\lambda \in [0, 1]$ and two actions $a_1, a_2 \in \mathbf{A}_D$, denote by $\lambda a_1 + (1-\lambda)\, a_2$ the lottery over elements of $\boldsymbol{\mathcal{R}}_D$ which associates with each $x \in X_D$ the set of outcome distributions $a_1\left(x\right)$ with probability λ and the set of outcome distributions $a_2\left(x\right)$ with probability $(1-\lambda)$. The resulting mass distribution is:

$$m_{\lambda a_1 + (1-\lambda) a_2}^f = \lambda m_{a_1}^f + (1-\lambda) m_{a_2}^f. \tag{2}$$

	$a \in \mathbf{A}_D$	$m_a^f \in \Delta\left(\boldsymbol{\mathcal{R}}_D\right)$		
	$(a(x_1), a(x_2))$	$m_a^f\left(\{\rho_1\}\right)$	$m_a^f\left(\{\rho_2\}\right)$	$m_a^f\left(\{\rho_1, \rho_2\}\right)$
a_1	$\left(\{\rho_1\}, \{\rho_1\}\right)$	1	0	0
a_2	$\left(\{\rho_1\}, \{\rho_2\}\right)$	f_1	f_2	0
a_3	$\left(\{\rho_1\}, \{\rho_1, \rho_2\}\right)$	f_1	0	f_2
a_4	$\left(\{\rho_2\}, \{\rho_1\}\right)$	f_2	f_1	0
a_5	$\left(\{\rho_2\}, \{\rho_2\}\right)$	0	1	0
a_6	$\left(\{\rho_2\}, \{\rho_1, \rho_2\}\right)$	0	f_1	f_2
a_7	$\left(\{\rho_1, \rho_2\}, \{\rho_1\}\right)$	f_2	0	f_1
a_8	$\left(\{\rho_1, \rho_2\}, \{\rho_2\}\right)$	0	f_2	f_1
a_9	$\left(\{\rho_1, \rho_2\}, \{\rho_1, \rho_2\}\right)$	0	0	1

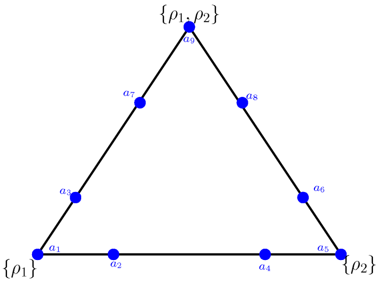


Table 3: Induced mass distributions: $m_a^f \in \Delta(\boldsymbol{\mathcal{R}}_D)$

Mixtures of acts in $\mathbf{A}_D$ are elements of the simplex $\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$. Given the frequency distribution f_D on X_D and the set of basic actions $\mathbf{A}_D$, let

$$\mathfrak{M}(\mathbf{A}_D, f_D) = \left\{ m_a^{f_D} = \sum_k \lambda_k m_{a_k}^{f_D} \mid a_k \in \mathbf{A}_D, k \in N \right\}$$

be the set of all mass distributions induced by mixed actions in $\mathbf{A}_D$.

The following Lemma 1 shows that the set of mass distributions $\mathfrak{M}(\mathbf{A}_D, f_D)$ obtained from all mixtures of actions in $\mathbf{A}_D$ equals the set of all mass distributions on $\boldsymbol{\mathcal{R}}_D$, $\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$, provided that $\mathbf{R}_D$ contains all Dirac measures over outcomes in R_D.

Lemma 1. $\mathfrak{M}\left(\mathbf{A}_D, f_D\right) = \Delta\left(\boldsymbol{\mathcal{R}}_D\right)$.

3.3 Actions on the extended set of characteristics

As argued in Section 2, a decision maker may also consider the possibility of characteristics x_o that have not been recorded in the data D. In this section, we extend the specification of actions to x_o.

Assume that a decision maker associates a finite set of possible outcome distribution $R_a^o \in \Re_D$ with x_o for every act $a \in A_D$. Allowing for $a\left(x_o\right) = R_a^o \in \Re_D$ amounts to assuming that the decision maker can in principle envision any hypothetical outcome distribution over observed outcomes.[14]

Given the extended set of characteristics $\hat{X}_D = X_D \cup \{x_o\}$, we consider actions in the extended action set $\mathbf{A}_D^o = \mathbf{A}_D \times \Re_D$. As shown in the previous subsection, each basic actions in $\mathbf{A}_D$ induces a mass distribution $m_a^{f_D} \in \Delta\left(\boldsymbol{\mathcal{R}}_D\right)$ which is then combined with the set of outcome distributions $R_a^o = a\left(x_o\right) \in \Re_D$ associated with x_o. Omitting the index f_D, an extended action a can be written as $(m_a, R_a) \in \Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$.

[14] We do not consider the possibility of ”new outcomes” and the resulting ”new actions” as in Karni and Vierø (2017), although such an extension is possible.

As in Subsection 3.2, we allow for mixtures on the extended set of actions $a \in \mathbf{A}^o_D$. Given two actions $a_1 = \left(m_{a_1}, R^o_{a_1}\right)$ and $a_2 = \left(m_{a_2}, R^o_{a_2}\right)$ and any $\lambda \in [0,1]$, define the convex combination of the two actions $\lambda a_1 + (1-\lambda) a_2$ as the action $a = \left(m_{\lambda a_1 + (1-\lambda) a_2}; R^o_{\lambda a_1 + (1-\lambda) a_2}\right)$ where $R^o_{\lambda a_1 + (1-\lambda) a_2}$ is the set of outcome distributions,

$$R^o_{\lambda a_1 + (1-\lambda) a_2} = \lambda R^o_{a_1} + (1-\lambda) R^o_{a_2} = \left\{\lambda \rho_1 + (1-\lambda) \rho_2 | \, \rho_1 \in R^o_{a_1}, \ \rho_2 \in R^o_{a_2}\right\}.$$

By construction, $m_{\lambda a_1 + (1-\lambda) a_2} \in \Delta(\boldsymbol{\mathcal{R}}_D)$ and $R^o_{\lambda a_1 + (1-\lambda) a_2} \in \Re_D$. That is, all convex combinations of extended actions $a = (m_a, R^o_a)$ will be elements of $\Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$.

We thus consider the set of actions $\mathcal{A}_D = \Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$. It is easy to check that $\mathcal{A}_D$ is a mixture set.

Lemma 2. *$\mathcal{A}_D = \Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$ is a mixture set.*

3.4 Preferences and Suggested Representation

Denote by $\succsim$ on $\Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$, the preference order of the decision maker on the set of actions $\mathcal{A}_D$. We assume that the decision maker can rank all actions in this set. Similarly to Savage (1954) this amounts to the ability to rank the consequences of the actions associated with the different characteristics. We will present axioms that imply that any action $a = (m_a, R^o_a) \in \mathcal{A}_D$ is evaluated by the following functional:

$$V_D(a) = (1-\gamma_D) \sum_{R \in \boldsymbol{\mathcal{R}}_D} m_a(R) \left[\alpha_D \max_{\rho \in R} \sum_r u(r) \rho(r) + (1-\alpha_D) \min_{\rho \in R} \sum_r u(r) \rho(r)\right] \tag{3}$$
$$+ \gamma_D \left[\alpha^o_D \max_{\rho \in R^o_a} \sum_r u(r) \rho(r) + (1-\alpha^o_D) \min_{\rho \in R^o_a} \sum_r u(r) \rho(r)\right]$$

where

- $u : R_D \to \mathbb{R}$ is a von Neumann-Morgenstern utility function over outcomes (unique up to a positive-affine transformation);

- $\alpha_D \in [0,1]$ is the degree of optimism and $(1-\alpha_D)$ is the degree of pessimism w.r.t. ambiguity in the outcome distributions given the observed data D, and $\alpha^o_D \in [0,1]$ and $(1-\alpha^o_D)$ are the degrees of optimism and pessimism for other, yet unobserved, characteristics. The degrees of optimism α_D and α^o_D may, but need not, coincide;

- $\gamma_D \in (0,1)$ is the degree of unawareness, the subjective weight assigned to x_o.

Any actually observed action $a \in A_D$ can be evaluated using the observed frequencies of characteristics $f_D(x)$ and the outcome predictions $\mathbf{R}_D(a,x)$ derived from the data:

$$V_D(a) = (1-\gamma_D) \sum_{x \in X_D} f_D(x) \left[\alpha_D \max_{\rho \in \mathbf{R}_D(a,x)} \sum_r u(r)\,\rho(r) + (1-\alpha_D) \min_{\rho \in \mathbf{R}(a,x)} \sum_r u(r)\,\rho(r) \right] \tag{4}$$

$$+ \gamma_D \left[\alpha_D^o \max_{\rho \in \mathbf{R}(a,x_o)} \sum_{r \in R_o} u(r)\,\rho(r) + (1-\alpha_D^o) \min_{\rho \in \mathbf{R}(a,x_o)} \sum_r u(r)\,\rho(r) \right]$$

As in state-based decision theory, different representations of preferences can be deduced from different systems of axiom, e.g., a smooth representation could be deduced from axioms as in Eichberger and Pasichnichenko (2021). The representation we choose has the advantage of having a small number of parameters, which can be easily estimated in experiments and can be used to study learning.

3.5 Axiomatization

We now provide an axiomatization of preferences over the set of actions $\mathcal{A}_D = \Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$, $a = (m_a, R_a^o)$. Our axiomatization builds on the approach in Jaffray (1989).

Axiom 1 The preference order $\succsim$ on $\mathcal{A}_D$ is complete, transitive and non-trivial in the following sense: there is an $R^o \in \boldsymbol{\mathcal{R}}_D$ and m_{a_1}, $m_{a_2} \in \Delta(\boldsymbol{\mathcal{R}}_D)$ such that for $a_1 = (m_{a_1}, R^o)$ and $a_2 = (m_{a_2}, R^o) \in \mathcal{A}_D$, $(m_{a_1}, R^o) \succ (m_{a_2}, R^o)$.

The non-triviality condition in Axiom 1 is somewhat stronger than usual. In particular, it requires that there is some outcome R^o associated with the "other characteristics" which is a subset of the actually observed probability distributions in the data and for which the decision maker is not fully indifferent among all extended actions. This non-triviality condition requires that the decision maker be not indifferent among all mass functions for at least some set of probability distributions associated with the "other" characteristics.

Axiom 2 For all $(m_{a_1}, R_{a_1}^o)$, $(m_{a_2}, R_{a_2}^o)$, $(m_{a_3}, R_{a_3}^o) \in \mathcal{A}_D$ and all $\lambda \in [0;1]$,

$$\begin{aligned}&(m_{a_1}, R_{a_1}^o) \succsim (m_{a_2}, R_{a_2}^o)\\ &\qquad \Leftrightarrow \lambda (m_{a_1}, R_{a_1}^o) + (1-\lambda)(m_{a_3}, R_{a_3}^o) \succsim \lambda (m_{a_2}, R_{a_2}^o) + (1-\lambda)(m_{a_3}, R_{a_3}^o).\end{aligned}$$

Axiom 3 For all $(m_{a_1}, R_{a_1}^o)$, $(m_{a_2}, R_{a_2}^o)$, $(m_{a_3}, R_{a_3}^o) \in \mathcal{A}_D$ such that $(m_{a_1}, R_{a_1}^o) \succ (m_{a_2}, R_{a_2}^o) \succ (m_{a_3}, R_{a_3}^o)$, there are $\lambda, \mu \in (0;1)$ such that

$$\lambda (m_{a_1}, R_{a_1}^o) + (1-\lambda)(m_{a_3}, R_{a_3}^o) \succ (m_{a_2}, R_{a_2}^o) \succ \mu (m_{a_1}, R_{a_1}^o) + (1-\mu)(m_{a_3}, R_{a_3}^o).$$

Remark 2. The three Axioms imply that preferences are separable across the two dimensions, m and R^o, see Proposition 2 in the Appendix.

The following corollary obtains:

Corollary 1. *Axioms 1–3 imply that for any $m_a, m_b \in \Delta(\boldsymbol{\mathcal{R}}_D)$, and any $R_a^o, R_b^o \in \Re_D$, $(m_a, R_a^o) \succsim (m_b, R_a^o)$ iff $(m_a, R_b^o) \succsim (m_b, R_b^o)$ and $(m_a, R_a^o) \succsim (m_a, R_b^o)$ iff $(m_b, R_a^o) \succsim (m_b, R_b^o)$.*

Remark 3. Note that in general m and R are different objects, m is a probability distribution on $\boldsymbol{\mathcal{R}}_D$, whereas R is an element of $\Re_D \supset \boldsymbol{\mathcal{R}}_D$. Nevertheless, each m which assigns a probability of 1 to a single set $R \in \boldsymbol{\mathcal{R}}_D$ can be uniquely identified with an $R^o \in \Re_D$ such that $R^o = R$. We could thus use the subset of actions for which the mass distributions have a singleton support, $\Delta(\boldsymbol{\mathcal{R}}_D)^C \times \Re_D$, as well as the subset of actions for which the set assigned to x_o is an element of $\boldsymbol{\mathcal{R}}_D$, $\Delta(\boldsymbol{\mathcal{R}}_D) \times \boldsymbol{\mathcal{R}}_D$ to formulate an Anscombe-Aumann-type axiom of state-independence of preferences. More specifically, to allow for different degrees of optimism in regard to observed and unobserved contingencies and thus for the possibility that two sets R and R' can be ranked when associated with the already observed or the unobserved characteristics. Thus, our state-independence axiom is imposed only on singleton sets.

Axiom 4 For any $\lambda \in [0, 1]$, any $\{\rho_1\}$, $\{\rho_2\}$ and $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$, and m and $m' \in \Delta(\boldsymbol{\mathcal{R}}_D)$ such that $m(\{\rho_1\}) = \lambda$, $m(\{\rho_2\}) = 1 - \lambda$, $m'(\{\rho\}) = 1$,

$$\begin{aligned} (\tilde{m}, \lambda\{\rho_1\} + (1-\lambda)\{\rho_2\}) &\succsim (\tilde{m}, \{\rho\}) \text{ for some } \tilde{m} \in \Delta(\boldsymbol{\mathcal{R}}_D) \text{ holds iff} \\ (m, R^o) &\succsim (m', R^o) \text{ holds for some } R^o \in \Re_D. \end{aligned}$$

While our last Axiom 4 concentrated on preferences with respect to singletons, we now turn to preferences regarding sets with multiple elements. Consider R and $R' \in \Re_D$. We will write $R \succsim^o R'$ iff $(m, R) \succsim (m, R')$ for some and thus, by Corollary 1, for all $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$. We will write $R \succsim^d R'$ iff for $m_a(R) = 1$ and $m_b(R') = 1$, $(m_a, R'') \succsim (m_b, R'')$ for some and thus, by Corollary 1, for all $R'' \in \Re_D$. Axiom 4 then implies that these two relations coincide for singletons $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$: $\{\rho\} \succsim^o \{\rho'\}$ iff $\{\rho\} \succsim^d \{\rho'\}$ for $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$, while for $\{\rho\}$ or $\{\rho'\} \in \Re_D \backslash \boldsymbol{\mathcal{R}}_D$, the preference $\succsim^d$ is not defined and the comparison of the two sets is determined by $\succsim^o$. In both cases, with a slight abuse of notation, we write $\rho \succsim \rho'$. Axiom 4 thus implies a well-defined preference order over the singleton sets, regardless of whether they are associated with observed or unobserved characteristics. This in turn allows us to define for each set of outcome distributions $R \in \Re_D$, a "best" and "worst" outcome distribution $\underline{\rho}_R, \overline{\rho}_R \in \Delta(R_D)$. The following axiom is an adaptation of the axiom introduced in Jaffray (1989):

Axiom 5 For all R, R', if $\underline{\rho}_R \succsim \underline{\rho}_{R'}$ and $\overline{\rho}_R \succsim \overline{\rho}_{R'}$, then $R \succsim^o R'$ and $R \succsim^d R'$.

Axiom 5 implies that the comparison between any two sets of outcome distributions only depends on their best and worst elements. In contrast to Jaffray (1989), however, preferences between sets of outcome distributions may depend on whether they are associated with an already observed characteristic, or a yet unobserved, "other" characteristic, i.e., $\succsim^d$ and $\succsim^o$ might differ on non-singleton sets.

Axioms 1–5 allow a representation similar to 3, but for the fact that the degrees of optimism in general depend on the best and worst outcomes in the set. To obtain α_D and α_D^o that are independent of the set, we impose two additional axioms, see Proposition 4 in the Appendix.

To understand the axiom, suppose that we compare two actions a and b with identical mass distributions $m_a = m_b =: m$. One of the actions attributes a set with two outcome distributions to x_o, $R_a = \{\rho, \rho'\}$ with $\rho \succ \rho'$, while, for the second action b, the set of outcome distributions on x_o contains only the mixture $R_b = \{\alpha\rho + (1-\alpha)\rho'\}$. Axioms 1–5 imply that there exists a unique $\alpha \in [0,1]$ such that $R_a \sim^o R_b$. This α is the weight assigned to the best outcome of the set R_a, i.e., the degree of optimism relative to unobserved characteristics with respect to this set. Axiom 6 postulates that the so-determined α is independent of the set R_a under consideration.

Axiom 6 For any ρ, ρ', ρ'', $\rho''' \in \Delta(R_D)$, such that $\rho \succ \rho'$ and $\rho'' \succ \rho'''$ and any $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$, let $R_a = \{\rho, \rho'\}$, $R_c = \{\rho'', \rho'''\}$ and for some $\alpha \in [0,1]$,

$$\begin{aligned} R_b &= \{\alpha\rho + (1-\alpha)\rho'\} \\ R_d &= \{\alpha\rho'' + (1-\alpha)\rho'''\} \end{aligned}$$

Then $(m, R_a) \sim (m, R_b)$ iff $(m, R_c) \sim (m, R_d)$.

The final Axiom is analogous to Axiom 6, but imposed on the set $\Delta(\boldsymbol{\mathcal{R}}_D)$, i.e., on the mass functions associated with the observed characteristics X_D. It implies that the degree of optimism relative to observed characteristics does not depend on the specific set under consideration.

Axiom 7 For any $\{\rho\}, \{\rho'\}, \{\rho''\}, \{\rho'''\} \in \boldsymbol{\mathcal{R}}_D$, such that $\rho \succ \rho'$ and $\rho'' \succ \rho'''$ and any $R \in \Re_D$, let[15] $m_a(\{\rho, \rho'\}) = 1$, $m_c(\{\rho'', \rho'''\}) = 1$ and for some $\alpha \in [0,1]$,

$$\begin{aligned} m_b(\{\rho\}) = \alpha,\ m_b(\{\rho'\}) = (1-\alpha) \\ m_d(\{\rho''\}) = \alpha,\ m_d(\{\rho'''\}) = (1-\alpha) \end{aligned}$$

Then $(m_a, R) \sim (m_b, R)$ iff $(m_c, R) \sim (m_d, R)$.

Axioms 1–7 are necessary and sufficient to obtain our desired representation:

Theorem 1. *The preference order $\succsim$ on $\mathcal{A}_D = \Delta(\boldsymbol{\mathcal{R}}_D) \times \Re_D$ satisfies Axioms 1–7, iff there is a representation*

$$V_D(m, R^o) = \sum_{R \in \boldsymbol{\mathcal{R}}_D} (1-\gamma_D)\, m(R) \left[\alpha_D \max_{\rho \in R} \sum_{r \in R_D} u(r)\,\rho(r) + (1-\alpha_D) \min_{\rho \in R} \sum_{r \in R_D} u(r)\,\rho(r)\right] \tag{5}$$

$$+ \gamma_D \left[\alpha_D^o \max_{\rho \in R^o} \sum_{r \in R_D} u(r)\,\rho(r) + (1-\alpha_D^o) \min_{\rho \in R^o} \sum_{r \in R_D} u(r)\,\rho(r)\right]$$

[15] Recall that $\boldsymbol{\mathcal{R}}_D$ is the set of all subsets of observed frequencies in the data, $\mathbf{R}_D$. Thus, if $\boldsymbol{\mathcal{R}}_D$ contains the singletons $\{\rho\}$, $\{\rho'\}$, $\{\rho''\}$ and $\{\rho'''\}$, then it also contains the sets $\{\rho, \rho'\}$ and $\{\rho'', \rho'''\}$ and vice-versa.

where u is a unique (up to a positive-affine transformation) von-Neumann-Morgenstern utility function over outcomes, $\gamma_D \in (0,1)$ is a unique parameter describing the perception of unawareness and $\alpha_D, \alpha_D^o \in [0;1]$ are unique parameters of optimism relevant to the set of observed, respectively, unobserved, characteristics.

A special case of the representation is the one in which the coefficient of optimism does not depend on the type of characteristics under consideration, i.e., $\alpha_D = \alpha_D^o$. Such a representation can be easily obtained by replacing Axiom 7 with the following one:

Axiom $7'$ For any $\{\rho\}, \{\rho'\} \in \boldsymbol{\mathcal{R}}_D$, such that $\rho \succ \rho'$ and any $R \in \Re_D$, let $m_a(\{\rho, \rho'\}) = 1$, and for some $\alpha \in [0,1]$,

$$m_b(\{\rho\}) = \alpha,\ m_b(\{\rho'\}) = (1-\alpha)$$

Then $(m_a, R) \sim (m_b, R)$ iff for some $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$,

$$(m, \{\rho, \rho'\}) \sim (m, \{\alpha\rho + (1-\alpha)\rho'\}). \tag{6}$$

Axiom $7'$ can be seen as an extension of Axiom 4 to sets in $\boldsymbol{\mathcal{R}}_D$ containing two elements. It requires that the mass function concentrated on a two element set $\{\rho, \rho'\}$is considered indifferent to a mass function which mixes the two singletons in proportions α (for the better one) and $(1-\alpha)$ for the worse one iff the same two element set, but assigned on the "other" characteristics is considered indifferent to the singleton set $\{\alpha\rho + (1-\alpha)\rho'\}$ which is a mixture of its two elements in the same proportions α and $(1-\alpha)$. Since by Axiom 6, the coefficient α does not depend on the choice of ρ and ρ', we obtain that the α_D^o identified in the proof of Theorem 1 applies also to all two-element sets $\{\rho'', \rho'''\} \in \boldsymbol{\mathcal{R}}_D \subset \Re_D$. We can thus set $\alpha_D^o = \alpha_D$ and obtain the following Corollary:

Corollary 2. *The preference order $\succsim$ on $\mathcal{A}_D$ satisfies Axioms 1–6 and $7'$, iff there is a representation*

$$V_D(m, R^o) = \sum_{R \in \boldsymbol{\mathcal{R}}_D} (1-\gamma_D)\, m(R) \left[\alpha_D \max_{\rho \in R} \sum_{r \in R_D} u(r)\rho(r) + (1-\alpha_D) \min_{\rho \in R} \sum_{r \in R_D} u(r)\rho(r)\right] \tag{7}$$

$$+\gamma_D \left[\alpha_D \max_{\rho \in R^o} \sum_{r \in R_D} u(r)\rho(r) + (1-\alpha_D) \min_{\rho \in R^o} \sum_{r \in R_D} u(r)\rho(r)\right]$$

where u is a unique (up to a positive-affine transformation) von-Neumann-Morgenstern utility function over outcomes, $\gamma_D \in (0,1)$ is a unique parameter describing the perception of unawareness and α_D is a unique parameter of optimism relevant both to the set of observed and the unobserved, characteristics.

4 Relevance of Cases and Identifying the Sets of Subjective Likelihoods

So far, we have taken as given the subjective likelihoods assigned to a given action-characteristic pair based on a data set D. This allowed us to extend preferences to the set of all mass distributions. In this section, we wish to identify the sets $R_D(a, x)$ directly from preferences.

To do so, we will make some assumptions on the class of data sets and the corresponding sets of actions, characteristics and outcomes. We will consider given sets of characteristics, X, actions, A and outcomes, R and the corresponding class of data sets $\mathbb{D}_X$ in which all observations satisfy $c \in X \times A \times R$. The support of a data set is defined as:

$$supp(D) = \{(a', x') \mid f_D(a', x') > 0\}$$

We start by extending the preference relation $\succsim_D$ to conditional preferences, i.e., to preferences over the choice of an action conditional on a subset of characteristics $\tilde{X}_D \subseteq \hat{X}_D$. The interpretation is that the decision maker can evaluate actions once the subset of possible characteristics is restricted to $\tilde{X}$. E.g., an algorithm may be evaluated based on a set of objects whose characteristics are included in $\tilde{X}$. Of particular interest will be conditional preferences on a single characteristic, $x \in \hat{X}_D$, $\succsim_{D,x}$. Given the additive structure of our representation, we can formally define $\succsim_{D,x}$as:

$$a \succsim_{D,x} a' \text{ iff } R_D(a, x) \succsim R_D(a', x)$$

For a given characteristic, x, the actions we consider associate with this characteristic, a set of probability distributions over $\Delta(R)$,

$$a_{|x} : \{x\} \rightrightarrows \Delta(R)$$

Formally, under our assumption that the decision maker entertains a set of subjective likelihoods for each action-characteristic pair, such actions represent a subset of the mass distributions in $\mathcal{A}_D$, namely constant actions, $\Delta(\mathcal{R}_D)^C$, given by a set $R \subseteq \Delta(R)$. However, differently from the mass distributions considered earlier, the set $R_D(a, x)$ is not specified and has to be uncovered from preferences. In this section, we will thus use $a_{|x} \in A$ for the actions observed in the data set used conditional on a characteristic x. As before, $R \subseteq \Delta(R)$ denotes a hypothetical constant action, i.e., a set of likelihoods. Conditional preferences $\succsim_{D,x}$ compare both types of actions, actual, as well as hypothetical. By definition, conditional preferences over hypothetical actions are identical to unconditional preferences $\succsim_D$ over constant actions:

$$R \succsim_{D,x} R'$$

iff

$$\begin{aligned} R &\succsim {}^d_D R' \text{ for } x \in X_D \\ R &\succsim {}^o_D R' \text{ for } x \in X_D \end{aligned}$$

Assuming that we can compare $a_{|x}$ to hypothetical constant actions, immediately implies that for each action $a_{|x}$ we can infer its certainty equivalent for the data set D, $R_{a|D,x}$ such that:

$$a_{|x} \sim_{D,x} R_{a|D,x}$$

However, this certainty equivalent is not unique and thus, the set of subjective likelihoods will not be uniquely identified.

As explained above, we will assume that the subjective likelihoods used to evaluate actions are drawn from among those observed in the data set. Formally,

AR1 (Experience-based beliefs) For each D, a, $x \in \hat{X}_D$, there is a certainty equivalent of $a_{|x}$ given by

$$R_D\left(a,x\right) \subseteq \left\{\rho_D\left(a',x'\right) \mid \left(a',x'\right) \in supp\left(D\right)\right\}$$

s.t.

$$R_D\left(a,x\right) \sim_{D,x} a_{|x}$$

Clearly, the set $R_D\left(a,x\right)$ is our candidate for the set of subjective likelihoods of outcomes given (a,x). We will write $S_{a,x}\left(D\right)$ for the set of action-characteristic pairs whose observed likelihoods enter the certainty equivalent:

$$S_{a,x}\left(D\right) = \left\{\left(a',x'\right) \in R_D\left(a,x\right)\right\}$$

We call the action-characteristic pairs in $S_{a,x}\left(D\right)$ the *most relevant for* (a,x) in the support of D. We now impose the following conditions on $S_{a,x}\left(D\right)$ and thus, implicitly on $R_{a|D,x}$:

We first postulate that relevance depends only on the support of the data set:

AR2 If $supp\left(D\right) = supp\left(D'\right) = S$, then $S_{a,x}\left(D\right) = S_{a,x}\left(D'\right)$ for all $(a,x) \in A \times X$.

According to axiom AR1, we can write $S_{a,x}\left(D\right) = S_{a,x}\left(supp\left(D\right)\right)$. Maximal relevance is thus a correspondence, which maps each support, $S \subseteq A \times X$ to a subset of S, $S_{a,x}\left(S\right) \subseteq S$. Mathematically, it is an object similar to a choice correspondence. We will thus impose the axioms of choice on the relevance correspondence $S_{a,x}\left(S\right) : 2^{A\times X} \rightarrow 2^{A\times X}$. Before doing so, we postulate that if present in the data, observations of (a,x) are the only most relevant observations for (a,x):

AR3 If $(a,x) \in S$, then $S_{a,x}\left(S\right) = \left\{(a,x)\right\}$.

Sen's axioms of choice are restated as:

AR4 If $\left(a',x'\right) \in S' \subseteq S''$ and if $\left(a',x'\right) \in S_{a,x}\left(S''\right)$, then $\left(a',x'\right) \in S_{a,x}\left(S'\right)$.

This is Sen's α-property, which states that if $\left(a',x'\right)$ is most relevant for (a,x) when the support is larger, then it is also most relevant when the support is smaller (provided $\left(a',x'\right)$ is still in the support).

AR5 If (a', x') and $(a'', x'') \in S_{a,x}(S')$, $S' \subseteq S''$ and $(a'', x'') \in S_{a,x}(S'')$, then $(a', x') \in S_{a,x}(S'')$.

This is Sen's β-property, which states that both (a', x') and (a'', x'') are most relevant for (a, x) in a given support, then expanding the support and finding that (a'', x'') is still most relevant implies that (a', x') should also be most relevant.

Taken together axioms AR1-AR5 imply

Proposition 1. *The class of preferences* $(\succsim_{D,x})_{\substack{x \in \hat{X}_D \\ D \in \mathbb{D}}}$ *satisfy axioms AR2-AR4 iff there exists for each* $(a, x) \in A \times X$ *a unique complete and transitive relevance order* $\geq_{a,x}$ *such that:*

- $(a', x') >_{a,x} (a'', x'')$ iff for some (and then all) S such that (a', x'), $(a'', x'') \in S$, $(a', x') \in S_{a,x}(S)$ and $(a'', x'') \notin S_{a,x}(S)$;
- $(a', x') =_{a,x} (a'', x'')$ iff for some (and then all) S such that (a', x'), $(a'', x'') \in S$, $(a', x') \in S_{a,x}(S)$ and $(a'', x'') \in S_{a,x}(S)$;
- $(a, x) >_{a,x} (a', x')$ for all $(a', x') \neq (a, x)$.

Furthermore, if preferences satisfy Axioms 1–7 as well as AR1-AR4, then the subjective likelihood sets $R_D(a, x)$ *can be written as:*

$$\begin{aligned} R_D(a, x) &= \{\rho_D(a', x') \mid (a', x') \in S_D(a, x)\} \\ &= \arg \max_{(a', x') \in supp(D)} \geq_{a,x} \end{aligned}$$

The proof follows from combining standard arguments in the theory of choice, see, e.g., Kreps (1988, p. 14) with Theorem 1, and is, therefore, omitted. The proposition provides an intuitive and simple rule for determining the subjective likelihoods, $R_D(a, x)$ for the actually observed actions a: the decision maker has a subjective relevance order, $\geq_{a,x}$ defined on the set of action-characteristic pairs, $X \times A$, which determines for each data set which of the available likelihoods are used for each action-characteristic pair, (a, x). A natural assumption is that (a, x) is a maximal (but potentially not unique) element of this order. The frequency of outcomes of (a', x'), $\rho_D(a', x')$, is used as a subjective likelihood for (a, x) if and only if (a', x') is the most relevant (but potentially not unique) w.r.t. $\geq_{a,x}$ pair observed in the data set.

Remark 4. The multiplicity of subjective likelihoods in the specification of $R_D(a, x)$ comes from the fact that when (a, x) is not observed in the data, multiple maximal relevant action-pair characteristics can be used to predict its outcome. The specification, however, does not take into account two additional sources of ambiguity: (i) ambiguity due to the limited number of observations for each action-characteristic pair; (ii) ambiguity due to the limited relevance of the action-characteristic pairs used for the prediction. A more general specification of $R_D(a, x)$ can be stated as:

$$R_D(a, x) = \{(1 - \epsilon_D(a', x'))\, \rho_D(a', x') + \epsilon_D(a', x')\, \delta_r \mid (a', x') \in S_{a,x}(D),\ r \in R\}$$

where $\epsilon_D(a', x') \in [0,1]$ is a coefficient of perceived ambiguity, which depends negatively both on the number of observations of (a', x') in D, as well as on the rank of (a', x') in the relevance order $\geq_{a,x}$. Note that when $\epsilon_D(a', x') = 0$ for all (a', x'), this specification coincides with the characterization in Proposition 1. For general values of ϵ, it replaces the actually observed likelihoods with their "discrete $\epsilon_D(a', x')$-contaminations", i.e., the set of likelihoods which results from scaling down the weight assigned to the observed frequencies to $(1 - \epsilon_D(a', x'))$ and assigning the remaining weight to one of the possible outcomes, r. Of particular interest for our representation are, of course, the best and the worst outcomes, since the corresponding subjective likelihoods will be actually used in the α-max-min evaluation of the action.

By replacing each observed relevant likelihood by a set of likelihoods, this representation captures the two types of ambiguity discussed above. To see this, consider first the case when D contains observations of (a, x). Then,

$$R_D(a,x) = \{(1 - \epsilon_D(a,x))\,\rho_D(a,x) + \epsilon_D(a,x)\,\delta_r \mid r \in R\}$$

The fact that ϵ is strictly decreasing in the number of observations implies that as more observations of (a, x) are observed, the ambiguity of the prediction will decrease and eventually converge to the actual frequency, $\rho_D(a, x)$. This captures the ambiguity due to limited number of observations which disappears when the data set becomes large.

If, instead, the data contain the same number of observations relevant for (a, x), but of a less relevant pair, (a', x'), the resulting ambiguity will incorporate both the ambiguity due to limited number of observations and that due to limited relevance and will be larger. As the number of observations increases, this ambiguity will decrease as well, but need not converge to 0.

These two effects were studied in Eichberger and Guerdjikova (2013). Using methods similar to those employed in this previous work, we can identify the corresponding coefficients of ambiguity and provide axioms for this more general characterization of subjective likelihoods.

Remark 5. Axiom AR3 can be relaxed to allow for multiple maximal elements of the order $\geq_{a,x}$ by requiring that (a, x) is always a maximal element, but not necessarily the unique one:

AR3′ If $(a, x) \in S$, then $(a, x) \in S_{a,x}(S)$.

A special case which could be of interest is the one in which $(a, x) = (a, x')$ for all x, x' which coincide on a given set of *relevant* categories, $\mathcal{T} \subset T$

$$x^t = x'^t \text{ for all } t \in \mathcal{T} \subset T$$

This corresponds to a coarsening of the set of characteristics by reducing the set of relevant categories. We write $x =_{\mathcal{T}} x'$ when x and x' coincide on all relevant characteristics. In this case, the relevance sets satisfy:

$$(a', x') \in S_{a,x}(S) \Longrightarrow (a', x'') \in S_{a,x}(S) \text{ for all } x'' \text{ s.t. } (a', x'') \in S \text{ and } x'' =_{\mathcal{T}} x'$$

The size of the set of characteristics that are differentiated by the decision maker is directly related to a measure of complexity, such as, for instance, the VC-dimension. Consider, e.g., the classification problem. Out of two decision makers such that $\succsim_D^1 = \succsim_D^2$ on the set of hypothetical actions (prescribed by hypothetical algorithms), but not necessarily on the actual actions. In particular, if for each x and D, we have that $S_x^1(D) \supseteq S_x^2(D)$, then decision maker 1 uses a coarser representation of the set of characteristics. This can be interpreted as $\succsim^1$ exhibiting stronger preferences for simplicity relative to $\succsim^2$.

5 Acquiring new data: The Classification Problem Revisited

In Section 3.5, we elicited from preferences, for a given data set, D the subjective parameters of the decision maker and in particular, the weight assigned on "other characteristics", x_o, γ_D, which can be interpreted as the degree of unawareness, and the parameters of optimism for the known characteristics, α_D and the "other" characteristics, α_D^o. In general, these three components of the representation depend on the available data set D.

Assume now that the decision maker obtains access to a new data set D'. This data set may take the form of a "continuation" of the history recorded in D, i.e., $D \subseteq D'$. For the purposes of the discussion below, we will use the classification problem described in Section 2.2.1 to illustrate the different scenarios. In this context, D may be the data set initially used for training, while D' corresponds to the training data set together with the correctly labeled test data-set, after testing has been completed and the correct classification revealed. We will write $D = \left\{ (x_n, r_n)_{n=1}^N \right\}$ and $D' = \left\{ (x'_n, r'_n)_{n=1}^{N'} \right\}$ for the two data sets in question. For the purposes of the following discussion, we will assume that the decision maker uses the same utility function u to evaluate lotteries over outcomes independently of the data set D.

5.1 Statistical learning

Statistical learning corresponds to the case in which no new characteristics are observed in the data. Instead, learning concerns the frequency of characteristics as well as the classification frequencies.

Consider a set of characteristics $X_D \subseteq X$ and an initial data set D. Consider a class of data sets which are a continuation of D and contain no new characteristics:

$$\mathbb{D}(D) = \left\{ D \circ \tilde{D} \mid X_{\tilde{D}} \subseteq X_D \right\}.$$

Relative to the initial data in D, the new data confirm that only characteristic in the set X_D are relevant. With $D' = D \circ \tilde{D}$, The decision maker would thus find it less likely that "other" characteristics could be observed in the future, suggesting $\gamma_{D'} < \gamma_D$ for each $D' \in \mathbb{D}(D)$. In a special, but particularly relevant case,

$D' = D \circ D$ exactly replicates the information in D, so that $f_{D'}(x) = f_D(x)$ and $\rho_D(a(k,x), x) = \rho_{D'}(a(k,x), x)$ remain unchanged for all x and all classification functions $a(k,x)$. By Proposition 1, the outcomes of "other" categories also remain unchanged, $\mathbf{R}_D(a_1, x_o) = \mathbf{R}_{D'}(a_1, x_o)$. Thus, the evaluations of algorithms conditional on the two data sets D and D' can only differ because of differences in the subjective parameters of unawareness γ and the degree of optimism, α^o. Since the new data confirm the already available evidence, the decision maker's attitude towards ambiguity should not change, $\alpha^o_{D'} = \alpha^o_D$.

For example, the following process for γ can be considered:

$$\gamma_{D'} = \frac{\gamma_{_D} N}{N + (1 - \gamma_{_D})(N' - N)} < \gamma_D.$$

One can think of the probability distribution resulting from D as determining the parameters of a Dirichlet distribution prior on $\hat{X}_D$ given by: $(1 - \gamma_D) f_D + \gamma_D \delta_{x_o}$. In this case, the formula above corresponds to the Bayesian update of the likelihood assigned to other characteristics based on the Dirichlet prior. Clearly, an increase in N' leads to a lower $\gamma_{D'}$, and thus, to a lower perception of unawareness. Using the number of instances of observation of "other" characteristics in the data set D, i.e., the total number of characteristics in X_D, $|X_D|$ as a proxy for γ_D allows us to further specify γ_D as:

$$\gamma_D = \frac{|X_D|}{N + |X_D|} \tag{8}$$

Finally, suppose that the set of characteristics X coincides with X_D, i.e., all characteristics have already been observed in the initial data set D. Suppose also that there is a distribution $\mu \in \Delta(X_D)$ such that the observed characteristic in a case i is drawn identically and independently from μ. For a given D, μ defines a measure π on the set of infinite sequences $(x_1, x_2...)$ observed in a data set in $\mathbb{D}(D)$. We then have that, on the class of data-sets $\mathbb{D}(D)$

$$\lim_{N \to \infty} f_D =_{\pi\text{-a.s.}} \mu$$

and since

$$\lim_{N \to \infty} \gamma_D = 0$$

$$\lim_{N \to \infty} (1 - \gamma_D) f_D + \gamma_D \delta_{x_o} =_{\pi\text{-a.s.}} \mu$$

We thus conclude that as the number of observations increase, the decision maker will use the actual probability distribution over characteristics μ and will assign a 0-weight to "other" characteristics leading to the evaluation of an action (π-a.s.) given by:

$$\lim_{N \to \infty} V_D(a) = \sum_{x \in X_D} \mu(x) \sum_k l_D(k,x) \ln(a(k,x)).$$

If, furthermore, the classification likelihoods also follow an i.i.d. process with probabilities $\lambda(k,x)$, $l_D(k,x)$, we will obtain π-a.s.,

$$\lim_{N \to \infty} V_D(a) = \sum_{x \in X_D} \mu(x) \sum_k \lambda(k,x) \ln(a(k,x)).$$

In the limit, evidence from data allows for learning whenever the set of characteristics is well identified and when the underlying uncertainty conforms to a stationary distribution. In this sense, our approach incorporates the standard Bayesian approach to statistical learning as a special case.

5.2 Learning new characteristics

As described above, statistical learning applies to well-structured environments in which new data "confirm" existing evidence. More generally, however, information in the form of data will make the decision maker aware of new features of the environment, such as the existence of new characteristics. We now turn to this scenario

5.2.1 Learning Other Characteristics

Consider first the case, in which the set of observed relevant categories T remains unchanged, but within a given category, a new characteristic is observed. Such a characteristic can correspond to a previously unobserved color of a ball drawn from an urn, or a new traffic situation to be learned by an autonomous vehicle (e.g., a low-flying wild turkey at risk to collide with the windshield of the car), etc.

Suppose thus that a data-set D' contains a single new characteristic previously unobserved in the data-set D, $x^{t,new}$ for category t, e.g.,

$$D' = D \circ \left(x^{new}, k\right)$$

How would the observation of new cases containing the new characteristic x^{new} change the decision maker's preferences? First, the observation $\left(x^{new}, k\right)$ will serve as confirmation for the existence of such "other" characteristics, x_o. This will reinforce the perceived ambiguity and increase γ so that $\gamma_{D'} > \gamma_D$. For instance, the specification of γ_D in (8) satisfies this property and increases with the observation of a new characteristic.

The second effect of the observation of the new characteristic consists in obtaining the empirical classification likelihood of an object with such "other" characteristics x_o, $l\left(x^{new}\right)$. This allows the decision maker to compare the "in-sample" performance of an algorithm (relative to the training data in D) to its "out-of-sample" performance in the test set. Let $\mathbf{B}^*\left(D\right)$ be the (set[16] of) optimal algorithms for data set D. We can now compare the in-sample performance of an optimal algorithm $\mathbf{a}^*\left(D\right) \in \mathbf{B}^*\left(D\right)$ on X_D to its "out-of-sample" performance on x^{new}:

$$V_{D'|X_D}\left(\mathbf{a}^*\left(D\right)\right) > (<)\, V_{D'|\{x^{new}\}}\left(\mathbf{a}^*\left(D\right)\right) \tag{9}$$

We will call the observation $\left(x^{new}, k\right)$ a positive surprise if

$$V_{D'|X_D}\left(\mathbf{a}^*\left(D\right)\right) < \min_{\mathbf{a}^*(D)\in\boldsymbol{B}^*(D)} V_{D'|\{x^{new}\}}\left(\mathbf{a}^*\left(D\right)\right) \tag{10}$$

[16] Non-trivial multiplicity can result for $a\left(k, x_o\right)$ even when the utility function u is strictly concave, because of the non-convexity of the optimization problem generated by a strictly positive degree of optimism.

and thus, all of the optimal algorithms for D perform strictly better out-of-sample than in sample. If instead,

$$V_{D'|X_D}\left(\mathbf{a}^*(D)\right) > \max_{\mathbf{a}^*(D)\in\mathbf{B}^*(D)} V_{D'|\{x^{new}\}}\left(\mathbf{a}^*(D)\right) \tag{11}$$

then all of the optimal algorithms for D perform strictly worse out-of-sample, we speak of a negative surprise. Intuitively, the observation of a new characteristic represents a positive surprise, when the decision maker learns that he is able to better forecast the class of "other" characteristics for any of the optimally determined algorithms than that of already observed ones. This may have an effect on the expectations the decision maker forms about his ability to predict the correct class for other yet unobserved characteristics making him more optimistic (pessimistic) increasing (decreasing) the degree of optimism α^o related to "other" characteristics, $\alpha^o_{D'} > (<)\, \alpha^o_D$.

Combining the learning about new characteristics with statistical learning allows for a dynamics similar to that discussed in Grant *et al.* (2017): suppose that the initial information provided to a decision-maker is a data-set within $\mathbb{D}(D)$. For such a data-set, the decision-maker entertains some degree of ambiguity γ_D. As long as incoming information confirms that characteristics come from the same set X_D, i.e., the data-set remains in $\mathbb{D}(D)$, perceived ambiguity decreases, while (absent observations of so far unlabeled characteristics), both the degree of optimism for observed and for unobserved characteristics remain constant. However, upon an observation of a new characteristic, the new data-set no longer belongs to $\mathbb{D}(D)$. The degree of ambiguity increases as a result of the experienced surprise. The new data set forms the base for a new class of data sets $\mathbb{D}(D')$. Afterwards, as long as no new characteristics are observed the decision maker reverts to statistical learning and ambiguity once again decreases.

5.2.2 Learning about Unlabeled Examples

We next consider the case in which the data set is augmented by an observation of a so far unlabeled characteristic, $\hat{x}$, $\left(\hat{x}, \hat{k}\right)$. Let $\mathbf{L}_D(\hat{x})$ be the initial prediction of the classification likelihood for the unlabeled example containing $\hat{x}$.

The labeled observation of $\hat{x}$ will have two effects: first, just as in the case of statistical learning, it will serve as confirmation that the relevant characteristics belong to the set X_D and will thus lead to a decrease in γ. However, just as in the case of observing a new characteristic, the new observation can serve as a test as to how well the optimal algorithms in D classify actual observations "out-of-sample" (for unlabeled examples). It is straightforward to restate the definition of a positive / negative surprise by replacing x^{new} in definitions (10) and (11) by $\hat{x}$ with the same interpretation as above.

A positive surprise with respect to unlabeled characteristics might have an effect on the expectations the decision maker forms about his ability to predict the correct class for characteristics with unobserved classification, making him more optimistic and increasing the degree of optimism α related to observed characteristics.

In Section 8 below, we state formally the axioms which imply the properties of the subjective perception of unawareness and degree of optimism discussed above in Sections 5.1, 5.2.1 and 5.2.2.

5.2.3 Contradictory evidence and new categories

When observable characteristics fail to uniquely predict the outcome of an action, the decision maker may perceive data as contradictory and suggest that some unobserved underlying factor may be relevant for the correct classification of objects. Consider, e.g., a data set D in which all objects with characteristic x are classified to be of class k, $l_D(x,k) = 1$. The decision maker, therefore, selects an algorithm which assigns predicts k for characteristic x with probability 1: $a(x,k) = 1$. Assume that this algorithm is applied to the test data set $\tilde{D}$ of the same length as D. The correct classification of the test data set is subsequently observed and results in $l_{\tilde{D}}(x,k') = 1$ with $k' \neq k$: in the test set, all objects with characteristic x are classified as k'. [17] How should the decision maker evaluate the algorithm (and in particular, $a(x,k)$) in the new data set $D' = D \circ \tilde{D}$? While it is certainly possible to use the combined frequency of observations, $l_{D'}(x,k) = l_{D'}(x,k') = \frac{1}{2}$, if the decision maker is cautious, when faced with the evidence he might be unwilling to commit himself to a specific likelihood, and instead, similarly to the approach suggested in Remark 4, use a set of likelihoods,

$$\mathbf{L}_{D'}(x) = \{\eta\delta_k + (1-\eta)\, l_{D'}(x), \eta\delta_{k'} + (1-\eta)\, l_{D'}(x)\}$$

This set combines the objective frequencies $l_{D'}(x,k)$ with the extreme scenarios of one of the observed classes having a probability of 1. Note that each such situation would correspond to the potential discovery of a characteristic for which the outcome x is classified as k (or k') for sure. Each of the distributions obtained is an "η-distance" away from the realized frequency. The parameter η can be interpreted as the subjective relevance assigned to the identification of a category which would allow to differentiate between the two classes to which x may belong. Special cases of such beliefs are $\eta = 0$, for which the relevance of identifying the unobserved category is null; and $\eta = 1$, for which it is maximal. Let $a^*(x)$ denote the classification prescribed by the optimal algorithm for x given η.

5.2.4 The value of new categories

We can now use the representation of preferences over algorithms to determine the value of identifying the new category. To do so, suppose that a data set $\tilde{D}'$ with a discriminatory category $X^{T+1} = \left\{x_1^{T+1}, x_2^{T+1}\right\}$ is discovered such that the correct class of $\left(x, x_1^{T+1}\right)$ is always k, whereas for $\left(x, x_2^{T+1}\right)$, it is k'. Let $f_{\tilde{D}'}\left(x, x_1^{T+1}, k\right) = f_{\tilde{D}'}\left(x, x_2^{T+1}, k'\right) = \frac{1}{2}$.

[17] Similar problems arise when randomized controlled trials are conducted in different countries and document varying levels of success of the policy intervention studied. A simple aggregation of the results without taking into account locally specific factors might significantly bias the results, see Deaton and Cartwright (2018).

Three effects obtain. First, the likelihoods for each characteristic, $\left(x, x_1^{T+1}\right)$ and $\left(x, x_2^{T+1}\right)$ become:

$$\begin{aligned} l_D\left(x, x_1^{T+1}\right) &= \delta_k \\ l_D\left(x, x_2^{T+1}\right) &= \delta_{k'} \end{aligned}$$

and the likelihood of an object with characteristic x (without the specification of x^{T+1}) to belong to class k or k' is $l_{\tilde{D}'}(x, k) = l_{\tilde{D}'}(x, k') = \frac{1}{2}$. This corresponds to a decrease in perceived ambiguity and will affect the evaluation of algorithms through the parameters of optimism and pessimism. For a pessimistic decision maker, $\alpha = 0$, the change in beliefs will positively affect the evaluation of all possible algorithms. For an optimistic decision maker ($\alpha = 1$), reduction of ambiguity will decrease the evaluation of the algorithms.

The second effect results from the fact that algorithms now can condition on the newly identified category. The optimal algorithm can now be adapted to predict $\tilde{a}^*\left(x, x_1^{T+1}, k\right) = 1$ and $\tilde{a}^*\left(x, x_2^{T+1}, k'\right) = 1$ to obtain a payoff of 1 in the training data set, regardless of the realization of x^{T+1}. This is Blackwell's effect, which is always positive.

The identification of a new category also expands the set of characteristics and, with it, the set of possible algorithms. This may correspond to an increase in complexity, as captured, e.g., by the VC-dimension of the set of possible algorithms, see Vapnik (2018, p. 145). The cost of such complexity is captured by the number of observations necessary to make reliable predictions. This, in turn, corresponds to the notion of ambiguity related to a limited number of observations. As long as the VC-dimension remains finite, a larger number of observations will be needed to obtain a prediction which limits the expected number of mistakes. This third effect may result in beliefs which are not single-valued, but take into account complexity is a source of ambiguity. In our example, the new subjective beliefs could be given by:

$$\begin{aligned} \mathbf{L}_{\tilde{D}'}\left(x, x_1^{T+1}\right) &= \left\{\delta_k, \left(1 - \epsilon_{\tilde{D}'}\left(x, x_1^{T+1}\right)\right)\delta_k + \epsilon_{\tilde{D}'}\left(x, x_1^{T+1}\right)\delta_{k'}\right\} \\ \mathbf{L}_{\tilde{D}'}\left(x, x_2^{T+1}\right) &= \left\{\delta_{k'}, \left(1 - \epsilon_{\tilde{D}'}\left(x, x_2^{T+1}\right)\right)\delta_{k'} + \epsilon_{\tilde{D}'}\left(x, x_1^{T+2}\right)\delta_k\right\} \end{aligned}$$

We call the difference in payoffs between the two optimal algorithms $a^*(x)$ and $\tilde{a}^*(x, \cdot)$

$$VNC(x, \eta) := V_{\tilde{D}'}\left(\tilde{a}^*(x, \cdot)\right) - V_{D'}\left(a^*(x)\right)$$

the value of identifying the new category X^{T+1}. VNC depends on the trade-off between the three effects described above and gives determine the willingness to pay for the discovery of new categories. Consider a pessimistic decision maker. If current ambiguity due to the contradictory nature of evidence is large, whereas the number of observations is sufficient, the negative effect of increase in complexity will be negligible, and the new category will be adopted. In contrast, if the amount of available data is small, the decision maker may group together categories to allow for faster learning, as discussed in Remark 5.

This discussion shows that our framework allows us to also formally model the perception of "other" categories, as well as the perception of "other" characteristics. This perception relies on beliefs about the nature and the structure of the data, which cannot however be inferred from the data set itself. Thus, just as with the perception of unawareness about other characteristics, the perception of unawareness about other categories is an individual subjective feature of the decision-maker and may depend on the context in which learning takes place. From a formal point of view, however, our representation derived in Theorem 1 can be used to evaluate actions given such unawareness and can also capture the fact that unawareness will disappear, once the relevant categories have been identified, potentially to be replaced by ambiguity due to the increase in complexity.[18]

6 Concluding remarks: From characteristics to states

In an ideal Savagean world, data are perfectly adapted to the description of uncertainty by a set of states of the world. In particular, the decision maker is aware of and knows all contingencies and there are no unobserved categories, or characteristics. Notably, the following three conditions are satisfied:

(i) The set X_D corresponds exactly to the Savagean state-space S, $X_D = S$.

(ii) Actions are functions from states to outcomes: $a : S \to R$. For each action, $a \in A_D$, and each characteristic, $x \in X_D$, exactly one outcome is observed in the data, i.e., $supp\,(\rho_D\,(\cdot \mid a, x))$ is a singleton for all a and x.

(iii) There are no redundant categories and characteristics, i.e., for each $t \in T$, $x^t \neq \tilde{x}^t$ implies that there is an $a \in A_D$ and $x^{-t} \in \Pi_{\tau \neq t} X_D^\tau$ such that $\rho_D\,(\cdot \mid a, (x^t, x^{-t})) \neq \rho_D\,(\cdot \mid a, (\tilde{x}^t, x^{-t}))$.

Suppose that the number of relevant contingencies is finite and that within a finite number of observations, all combinations (a, x) are observed in the data. If conditions *(i)*–*(iii)* are satisfied, arrival of new data corresponds to the case of statistical learning as described in Section 5.1. Since all possible characteristics have already been observed, no surprises occur. Thus, as the number of observations goes to ∞, $\gamma_D \to 0$. The decision maker behaves like an expected utility maximize w.r.t. the state space $S = X_D$. Probabilities coincide with limit frequencies recorded in the data.

Few decision situations satisfy the conditions listed above. A decision maker who wishes to be a Savagean, but is faced with empirical data that do not perfectly fit the desiderata has to learn the best approximation of such a model given available evidence.

[18] Grant *et al.* (2020) model the perception of ambiguity due to unawareness of propositions, as well as the process by which such ambiguity diminishes as the decision-maker's awareness increases.

The first type of learning was discussed in Section 5.2 and concerns becoming aware of new characteristics. We can model awareness of such unawareness by using a placeholder characteristic x_o which is taken into account for the evaluation of actions. As explained above, such learning increases the set of relevant characteristics, while at the same time increasing the degree of unawareness concerning the existence of "other" yet unobserved characteristics.

The second type of learning concerns the learning of new categories discussed in Section 5.2.3. For a Savagean decision maker, an action which results in two distinct outcomes for a given state entails a contradiction and signals that the state-space is not well-specified. Call a data set D consistent if for each a and x, a has resulted in a single outcome in combination with x and thus, $supp(\rho_D(\cdot \mid a, x))$ is a singleton. Otherwise, we call the data set inconsistent. Let

$$X_D^C = \{x \mid supp(\rho_D(\cdot \mid a, x)) \text{ is not a singleton for some } a\}$$

be the set of characteristics for which such indeterminacy of outcomes has been generated. From the point of view of a Savagean decision maker, these are the characteristics in need of refinement if they were to represent states of the world. The existence of such characteristics signals the decision maker's awareness that he is unaware of some relevant categories, without knowing explicitly what those could be. Such awareness of unawareness, in a natural way, also leads to ambiguity: the decision maker assigns multiple payoffs to an already observed (a, x) combination, rather than using the generated frequency of outcomes in the data.

In turn, observing a relevant category t^{new} such that $supp(\rho_D(\cdot \mid a, (x, x^{new})))$ is a singleton for each realization of the characteristic x^{new}, restores consistency of the data set, but might generate ambiguity if the relevant characteristics have not been measured for past observations.

As the decision maker learns new categories and thus, the elements of the support of $\rho_D(\cdot \mid a, x)$ can be attributed to distinct characteristics, and as sufficient observations of the so-refined characteristics are gathered, so that the number of missing observations becomes negligible, the two types of ambiguity related to categories also vanish.

Finally, new measurement methods might lead to the observation of new categories, even though these might appear redundant given the empirical information available. In particular, if $\rho_D(\cdot \mid a, (x^t, x^{-t})) = \rho_D(\cdot \mid a, (\tilde{x}^t, x^{-t}))$ holds for all x^t, $\tilde{x}^t \in X^t$, all $x^{-t} \in \Pi_{\tau \neq t} X_D^\tau$ and all $a \in A_D$, the decision maker may decide that the relevant state-space $S = \Pi_{\tau \neq t} X_D^\tau$ is a sufficient description of the underlying uncertainty, all be it coarser than the one suggested by the data set, X_D.

Whether or not such coarsening of the state-space is warranted will be an empirical question. As data accumulate, such coarsening might need to be reversed, as new observations might result in an inconsistent data set signaling that category t is not redundant after all. The process described above would then repeat.

The preceding discussion is closely related to the literature on unawareness. Notably, the type of learning described in Section 4.2 corresponds to the decision maker initially perceiving a reduction of the actual state-space, which then expands to take into account new contingencies, see e.g., Grant and Quiggin (2013a,b);

Grant *et al.* (2017). In contrast, learning new categories (Section 4.3) models an initial situation of coarsening of the state-space, which is sequentially refined, see, e.g., Grant and Quiggin (2006); Dominiak and Guerdjikova (2021). The works of Karni and Vierø (2013); Karni and Vierø (2017) and Vierø (2021) also model expansion of the state-space, though one related to the discovery of new acts or new outcomes, as opposed to learning new characteristics. The model we presented here provides a unified framework which captures these phenomena and relates them to empirical data.

Finally, the issue of complexity and related to it, the ability to learn the actual distribution of states and outcomes, is highly relevant. Refinements of the set of characteristics increase the complexity of the problem (formally, the VC-dimension) so that learning requires a larger set of observations. Modelling the cost of complexity and the related problem of structural risk-minimization are left as avenue for future research.

7 Appendix: Proofs

We prove the results of Theorem 1 using a sequence of Propositions. Combining the result of Jaffray (1989) with the implications of the first three axioms in the Anscombe-Aumann framework, see Kreps (1988, p. 102), we obtain:

Proposition 2. *Preferences $\succsim$ on $\mathcal{A}_D$ satisfy Axioms 1 – 3 iff there exist functions $U : \boldsymbol{\mathcal{R}}_D \to \mathbb{R}$ and $U_o : \Re_D \to \mathbb{R}$ such that for $a = (m_a, R^o_a)$ and $b = (m_b, R^o_b)$*

$$\begin{array}{rcl} (m_a, R^o_a) & \succsim & (m_b, R^o_b) \text{ iff} \\ \sum\limits_{R \in supp(m_a)} m_a(R)\, U(R) + U_o(R^o_a) & \geq & \sum\limits_{R \in supp(m_b)} m_b(R)\, U(R) + U_o(R^o_b) . \end{array}$$

Furthermore, U_o is affine and U and U_o are unique up to a positive-affine transformation with a common multiplication factor $z_1 > 0$.

Lemma 3. *Assume that preferences $\succsim$ satisfy Axioms 1–4.*

(i) For some singleton sets $\{\bar{\rho}\}$ and $\{\underline{\rho}\} \in \boldsymbol{\mathcal{R}}_D$ and for any $\tilde{m} \in \Delta(\boldsymbol{\mathcal{R}}_D)$,

$$(\tilde{m}, \{\bar{\rho}\}) \succsim (\tilde{m}, \{\rho\}) \succsim (\tilde{m}, \{\underline{\rho}\}) \tag{12}$$

holds for all singleton sets $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$.

(ii) For $\bar{m}$ and $\underline{m}$ satisfying $\bar{m}(\{\bar{\rho}\}) = 1$ and $\underline{m}(\{\underline{\rho}\}) = 1$, and for any $R^o \in \Re_D$,

$$(\bar{m}, \{R^o\}) \succsim (m, R^o) \succsim (\underline{m}, R^o),$$

holds for any m with $m(\{\rho\}) = 1$ for some singleton $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$.

(iii) If

$$(\bar{m}, \{\bar{\rho}\}) \succ (\underline{m}, \{\underline{\rho}\})$$

there exists a unique $\gamma_D \in [0, 1]$ such that

$$(\underline{m}, \{\bar{\rho}\}) \sim (\gamma_D \bar{m} + (1 - \gamma_D)\, \underline{m}, \gamma_D \{\bar{\rho}\} + (1 - \gamma_D) \{\underline{\rho}\}) .$$

Using Axioms 1–4 we obtain the following result.

Proposition 3. *Suppose that $\succsim$ satisfy Axioms 1–4. If*

$$(\bar{m}, \{\bar{\rho}\}) \succ (\underline{m}, \{\bar{\rho}\}) \succ (\underline{m}, \{\underline{\rho}\}), \tag{13}$$

then γ_D satisfies $\gamma_D \in (0,1)$. Furthermore, there exist functions $U : \boldsymbol{\mathcal{R}}_D \to \mathbb{R}$ and $U_O : \Re_D \to \mathbb{R}$ such that for $a = (m_a, R_a^o)$ and $b = (m_b, R_b^o)$

$$\begin{aligned} (m_a, R_a^o) &\succsim (m_b, R_b^o) \textit{ iff} \\ &\quad (1-\gamma_D) \sum_{R \in supp(m_a)} m_a(R) U(R) + \gamma_D U_O(R_a^o) \\ &\geq (1-\gamma_D) \sum_{R \in supp(m_b)} m_b(R) U(R) + \gamma_D U_O(R_b^o), \end{aligned}$$

where U is the function identified in Proposition 2 and there is a unique γ_D such that $U_O(\{\rho\}) = U(\{\rho\})$ for all singleton sets $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$ and $U_O(R) = \frac{1-\gamma_D}{\gamma_D} U_o(R)$ for any $R \in \Re_D$.

U_O is affine and U and U_O are unique up to a positive-affine transformation with a common multiplication factor $z_1 > 0$.

A consequence of the last part of the proof of Proposition 3 is that $\{\bar{\rho}\}$ and $\{\underline{\rho}\}$ are also the best and the worst singleton elements of $\Re_D$ and that each of them can be taken to be a Dirac measure on a single outcome, $\bar{\rho} = \delta_{\bar{r}}$, $\underline{\rho} = \delta_{\underline{r}}$, where $\bar{r}$ and $\underline{r}$, are respectively the best and the worst outcome in R_D.

Corollary 3. *The two inequalities in (12) hold for all singleton sets $\{\rho\} \in \Re_D$. Furthermore, one can set $\underline{\rho} = \delta_{\underline{r}}$ and $\bar{\rho} = \delta_{\bar{r}}$, for some $\underline{r}$ and $\bar{r} \in R_D$. Finally, there exists a utility function over outcomes $u : R_D \to \mathbb{R}$ which is unique up to a positive-affine transformation and satisfies:*

$$u(r) = U(\{\delta_r\}) = U_O(\{\delta_r\})$$

for any $r \in R_D$ and

$$\begin{aligned} U(\{\rho\}) &= \sum_{r \in R_D} u(r) \rho(r) \textit{ for all } \{\rho\} \in \boldsymbol{\mathcal{R}}_D \\ U_O(\{\rho\}) &= \sum_{r \in R_D} u(r) \rho(r) \textit{ for all } \{\rho\} \in \Re_D. \end{aligned}$$

A straightforward adaptation of Jaffray (1989)'s result yields the following proposition.

Proposition 4. *The preference order $\succsim$ on $\mathcal{A}_D$ satisfies Axioms 1–5 iff there exist a non-constant functions $w^d : \Delta(R_D) \times \Delta(R_D) \to \mathbb{R}$ non-decreasing w.r.t. the order $\succsim^d$ in its arguments, a non-constant function $w^o : \Delta(R_D) \times \Delta(R_D) \to \mathbb{R}$*

non-decreasing w.r.t. the order $\succsim^o$ *in its arguments, and a unique weight* $\gamma_D \in (0,1)$ *such that*

$$\begin{aligned}(m_a; R_a) &\succsim (m_b; R_b) \text{ iff} \qquad (14)\\ &\sum_{R\in\mathcal{R}_D} (1-\gamma_D)\, m_a(R)\, w^d\left(\underline{\rho}_R, \bar{\rho}_R\right) + \gamma_D w^o\left(\underline{\rho}_{R_a}, \bar{\rho}_{R_a}\right)\\ \geq &\sum_{R\in\mathcal{R}_D} (1-\gamma_D)\, m_b(R)\, w^d\left(\underline{\rho}_R, \bar{\rho}_R\right) + \gamma_D w^o\left(\underline{\rho}_{R_b}, \bar{\rho}_{R_b}\right).\end{aligned}$$

where $w^d(\rho,\rho) = w^o(\rho,\rho)$ *for all* ρ *such that* $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$ *and* $w^o(\rho,\rho) = \sum_{r\in R_D} u(r)\,\rho(r)$*, where* u *is the utility function over outcomes. The weights* w^o *and* w^d *are unique up to a positive-affine transformation with a common factor* $z_1 > 0$*.*

<u>Proof of Theorem 1</u>

By representation (14), we can write for any $R \in \Re_D$,

$$U_O(R) = \alpha^o_D\left(\underline{\rho}_R, \bar{\rho}_R\right) w^o(\bar{\rho}_R, \bar{\rho}_R) + \left(1-\alpha^o_D\left(\underline{\rho}_R, \bar{\rho}_R\right)\right) w^o\left(\underline{\rho}_R, \underline{\rho}_R\right) = U_O\left(\left\{\underline{\rho}_R, \bar{\rho}_R\right\}\right) \qquad (15)$$

and for any $R \in \boldsymbol{\mathcal{R}}_D$,

$$U(R) = \alpha_D\left(\underline{\rho}_R, \bar{\rho}_R\right) w^o(\bar{\rho}_R, \bar{\rho}_R) + \left(1-\alpha_D\left(\underline{\rho}_R, \bar{\rho}_R\right)\right) w^o\left(\underline{\rho}_R, \underline{\rho}_R\right) = U\left(\left\{\underline{\rho}_R, \bar{\rho}_R\right\}\right) \qquad (16)$$

It is thus sufficient to determine U_O and U for sets with two elements (the case $\underline{\rho}_R = \bar{\rho}_R$ has already been discussed above).

Consider thus ρ, $\rho' \in \Delta(R_D)$ with $\rho \succ \rho'$, the corresponding set of these two outcome distributions $R_a = \{\rho, \rho'\}$ and the singleton set $R_b = \{\alpha\rho + (1-\alpha)\rho'\}$ for some $\alpha \in [0,1]$. By continuity, Axiom 3, there exists a unique α such that for some (and then all) $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$, $(m, R_a) \sim (m, R_b)$. By Proposition 4, we then have:

$$\begin{aligned}U_O(R_a) &= \alpha(\rho',\rho) \sum_{r\in R_D} u(r)\,\rho(r) + (1-\alpha(\rho',\rho)) \sum_{r\in R_D} u(r)\,\rho'(r)\\ &= \alpha \sum_{r\in R_D} u(r)\,\rho(r) + (1-\alpha) \sum_{r\in R_D} u(r)\,\rho'(r) = U_O(R_b)\end{aligned}$$

and, thus, $\alpha^o_D(\rho',\rho) = \alpha$.

By Axiom 6, for any ρ'', $\rho''' \in \Delta(R_D)$ with $\rho'' \succ \rho'''$, the corresponding set of these two outcome distributions $R_c = \{\rho'', \rho'''\}$ and the singleton set $R_d = \{\alpha\rho'' + (1-\alpha)\rho'''\}$ we have $(m, R_c) \sim (m, R_d)$. We thus obtain $\alpha^o_D(\rho''', \rho'') = \alpha^o_D(\rho', \rho) = \alpha$ for any ρ, ρ', ρ'' and ρ'''. It follows that

$$\alpha^o_D\left(\underline{\rho}_R, \bar{\rho}_R\right) = \alpha$$

for all $\underline{\rho}_R$ and $\bar{\rho}_R$ and thus, for all $R \in \Re_D$. Setting $\alpha^o_D = \alpha$ thus implies that for any $R \in \Re_D$,

$$U_O(R) = \alpha^o_D \max_{\rho\in R} \sum_{r\in R_D} u(r)\,\rho(r) + (1-\alpha^o_D) \min_{\rho\in R} \sum_{r\in R_D} u(r)\,\rho(r) \qquad (17)$$

and thus, the optimism parameter for unobserved characteristics x_o does not depend on the set of outcomes R.

The argument that Axiom 7 implies that there exists an $\alpha_D \in [0,1]$ such that for any $R \in \boldsymbol{\mathcal{R}}_D$,

$$U(R) = \alpha_D \max_{\rho \in R} \sum_{r \in R_D} u(r)\rho(r) + (1-\alpha_D) \min_{\rho \in R} \sum_{r \in R_D} u(r)\rho(r) \tag{18}$$

is analogous and thus omitted.

Combining the representation in Proposition 3 with the expressions in (17) and (18) gives the desired representation.□

7.1 Online Appendix

Proof of Lemma 1:

Note that for each set of outcome distributions $R \in \Re_D$, there is a constant action $\bar{a}_R$ which has R as the set of outcomes for each $x \in X_D$, $a(x) \equiv R$. The Dirac measures $\delta_R \in \Delta(\boldsymbol{\mathcal{R}}_D)$ are elements of $\mathfrak{M}(\mathbf{A}_D, f_D)$ since for each $R \in \Re_D$, $\bar{a}_R \in \mathbf{A}_D$.

Moreover, the set of Dirac measures $\{\delta_R \mid R \in \Re_D\}$ are the extreme points of the simplex $\Delta(\boldsymbol{\mathcal{R}}_D)$. By Carathéodory's theorem, every $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$ can be obtained as a convex combination of these extreme points. □

Proof of Lemma 3:

(i) Note that according to Corollary 1, by the finiteness of $\boldsymbol{\mathcal{R}}_D$ and the fact that $\{\delta_r\} \in \boldsymbol{\mathcal{R}}_D$ for each $r \in R_D$, we have that there exist singleton sets $\{\bar{\rho}\}$ and $\{\underline{\rho}\} \in \boldsymbol{\mathcal{R}}_D$ such that for some (and thus, for any) $\tilde{m} \in \Delta(\boldsymbol{\mathcal{R}}_D)$,

$$(\tilde{m}, \{\bar{\rho}\}) \succsim (\tilde{m}, \{\rho\}) \succsim (\tilde{m}, \{\underline{\rho}\})$$

holds for all singleton sets $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$. We refer to $\{\bar{\rho}\}$ and $\{\underline{\rho}\}$ as the best and the worst singleton element of $\boldsymbol{\mathcal{R}}_D$.

(ii) Given the statement of part (i), the implication of Axiom 4 is that for $\bar{m}$ and $\underline{m}$ satisfying $\bar{m}(\{\bar{\rho}\}) = 1$ and $\underline{m}(\{\underline{\rho}\}) = 1$, and any m with $m(\{\rho\}) = 1$ for some singleton $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$, we have for some (and thus for any) $R^o \in \Re_D$,

$$(\bar{m}, \{R^o\}) \succsim (m, R^o) \succsim (\underline{m}, R^o),$$

i.e., $\bar{m}$ and $\underline{m}$ are the best and the worst elements in $\Delta(\boldsymbol{\mathcal{R}}_D)$ among those assigning full mass to singleton sets.

(iii) To show part (iii), consider next the extended action defined by $(\underline{m}, \{\bar{\rho}\})$. We have, by Corollary 1,

$$(\bar{m}, \{\bar{\rho}\}) \succsim (\underline{m}, \{\bar{\rho}\}) \succsim (\underline{m}, \{\underline{\rho}\}).$$

If $(\bar{m}, \{\bar{\rho}\}) \succ (\underline{m}, \{\underline{\rho}\})$, by continuity, Axiom 3, there exists a unique $\gamma_D \in [0,1]$ such that

$$(\underline{m}, \{\bar{\rho}\}) \sim (\gamma_D \bar{m} + (1-\gamma_D)\underline{m}, \gamma_D \{\bar{\rho}\} + (1-\gamma_D)\{\underline{\rho}\})$$

<u>Proof of Proposition 3</u>:

Using γ_D identified in part *(iii)* of Lemma 3 and the representation from Proposition 2, we have:

$$\begin{aligned} U\left(\{\underline{\rho}\}\right)+U_o\left(\{\bar{\rho}\}\right) &= \gamma_D U\left(\{\bar{\rho}\}\right)+\left(1-\gamma_D\right)U\left(\{\underline{\rho}\}\right) \\ &\quad +\gamma_D U_o\left\{\bar{\rho}\right\}+\left(1-\gamma_D\right)U_o\left\{\underline{\rho}\right\} \end{aligned}$$

$$\left(1-\gamma_D\right)\left[U_o\left(\{\bar{\rho}\}\right)-U_o\left(\{\underline{\rho}\}\right)\right]=\gamma_D\left[U\left(\{\bar{\rho}\}\right)-U\left(\{\underline{\rho}\}\right)\right] \tag{19}$$

When condition (13) holds, we have that $\gamma_D \notin \{0,1\}$ and we can rewrite (19) as:

$$\left[U_o\left(\{\bar{\rho}\}\right)-U_o\left(\{\underline{\rho}\}\right)\right]=\frac{\gamma_D}{1-\gamma_D}\left[U\left(\{\bar{\rho}\}\right)-U\left(\{\underline{\rho}\}\right)\right] \tag{20}$$

By continuity, Axiom 3, we have that for any singleton $\{\rho\}\in\boldsymbol{\mathcal{R}}_D$, there is a unique coefficient $\lambda_\rho\in[0,1]$ such that for any $\tilde{m}\in\Delta\left(\boldsymbol{\mathcal{R}}_D\right)$

$$\left(\tilde{m},\lambda_\rho\left\{\bar{\rho}\right\}+\left(1-\lambda_\rho\right)\left\{\underline{\rho}\right\}\right)\sim\left(\tilde{m},\{\rho\}\right)$$

and by Axiom 4, this is equivalent to the statement that for m such that $m\left(\{\rho\}\right)=1$ and any $R^o\in\Re_D$,

$$\left(\lambda_\rho\bar{m}+\left(1-\lambda_\rho\right)\underline{m},R^o\right)\sim\left(m,\{R^o\}\right).$$

Hence, normalizing, w.l.o.g. $U\left(\{\underline{\rho}\}\right)=U_o\left(\{\underline{\rho}\}\right)=0$ and $U\left(\{\bar{\rho}\}\right)=1$ and thus, by (20) $U_o\left(\{\bar{\rho}\}\right)=\frac{\gamma_D}{1-\gamma_D}$, we obtain that for any $\{\rho\}\in\boldsymbol{\mathcal{R}}_D$,

$$\begin{aligned} U\left(\{\rho\}\right) &= \lambda_\rho \\ U_o\left(\{\rho\}\right) &= \frac{\gamma_D}{1-\gamma_D}\lambda_\rho. \end{aligned}$$

Using the representation in Proposition 2, we thus obtain that for m_a and m_b which put their entire mass on singleton sets, i.e., $supp\left(m_a\right)$, $supp\left(m_b\right)\subseteq\left\{\{\rho\}\in\boldsymbol{\mathcal{R}}_D\right\}$ and $R_a^o=\{\rho_a\}$, $R_b^o=\{\rho_b\}$ for some singletons $\{\rho_a\}$, $\{\rho_b\}\in\boldsymbol{\mathcal{R}}_D$, we have

$$\begin{aligned} \left(m_a,R_a^o\right) &\succsim \left(m_b,R_b^o\right)\text{ iff} \\ &\quad \sum_{\{\rho\}\in supp(m_a)} m_a\left(\{\rho\}\right)U\left(\{\rho\}\right)+\frac{\gamma_D}{1-\gamma_D}U\left(\{\rho_a\}\right) \\ &\geq \sum_{\{\rho\}\in supp(m_b)} m_b\left(\{\rho\}\right)U\left(\{\rho\}\right)+\frac{\gamma_D}{1-\gamma_D}U\left(\{\rho_b\}\right) \\ &\quad\text{iff} \end{aligned}$$

$$\begin{aligned} &\left(1-\gamma_D\right)\sum_{\{\rho\}\in supp(m_a)} m_a\left(\{\rho\}\right)U\left(\{\rho\}\right)+\gamma_D U\left(\{\rho_a\}\right) \\ \geq\ &\left(1-\gamma_D\right)\sum_{\{\rho\}\in supp(m_b)} m_b\left(\{\rho\}\right)U\left(\{\rho\}\right)+\gamma_D U\left(\{\rho_b\}\right). \end{aligned}$$

Note further than since U_o is affine, we have that for any $\rho \in \Delta(R_D)$,

$$U_o(\{\rho\}) = \sum_{r \in R_D} \rho(r) U_o(\{\delta_r\}).$$

Since $\{\delta_r\} \in \boldsymbol{\mathcal{R}}_D$ for every $r \in R_D$, this implies, that we can set $\underline{\rho} = \delta_{\underline{r}}$ and $\bar{\rho} = \delta_{\bar{r}}$, where $\underline{r}$ is the "worst" and $\bar{r}$, the "best" outcome in R_D. It is then obvious that the two inequalities in (12) hold for all singleton sets $\{\rho\} \in \Re_D$, and we can thus refer to $\{\bar{\rho} = \delta_{\bar{r}}\}$ and $\left\{\underline{\rho} = \delta_{\underline{r}}\right\}$ as the best and the worst singleton element of $\Re_D$. Thus, we can define $u(r) = U_O(\{\delta_r\}) = \frac{1-\gamma_D}{\gamma_D} U_o(\{\delta_r\}) = U(\{\delta_r\})$ for every $r \in R_D$ so as to obtain for any $\rho \in \Delta(R_D)$

$$U_o(\{\rho\}) = \frac{\gamma_D}{1-\gamma_D} \sum_{r \in R_D} \rho(r) U(\{\delta_r\}) = \frac{\gamma_D}{1-\gamma_D} \sum_{r \in R_D} \rho(r) u(r) \tag{21}$$

and any $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$

$$U(\{\rho\}) = \sum_{r \in R_D} \rho(r) U(\{\delta_r\}).$$

Define the function $U_O : \Re_D \to \mathbb{R}$ as follows. Let

$$U_O(\{\rho\}) = U(\{\rho\}) = \sum_{r \in R_D} \rho(r) U(\{\delta_r\}) = \sum_{r \in R_D} \rho(r) u(r) \tag{22}$$

for all $\rho \in \Delta(R_D)$. Provided that $\gamma_D \neq 0$, for any $R \in \Re_D$, we can define $U_O(R) = \frac{1-\gamma_D}{\gamma_D} U_o(R)$ (note that by (21) and (21), this equality also holds for singletons $\{\rho\} \in \Re_D$). The so-defined $U_O(R) = \frac{1-\gamma_D}{\gamma_D} U_o(R)$ is a positive-affine transformation of U_o determined in Proposition 2. Indeed, we obtain that for $a = (m_a, R_a^o)$ and $b = (m_b, R_b^o)$

$$\begin{aligned} (m_a, R_a^o) &\succsim (m_b, R_b^o) \text{ iff} \\ & \sum_{R \in supp(m_a)} m_a(R) U(R) + U_o(R_a^o) \\ &\geq \sum_{R \in supp(m_b)} m_b(R) U(R) + U_o(R_b^o) \end{aligned}$$

iff

$$\begin{aligned} & \sum_{R \in supp(m_a)} m_a(R) U(R) + \frac{\gamma_D}{1-\gamma_D} U_O(R_a^o) \\ \geq & \sum_{R \in supp(m_b)} m_b(R) U(R) + \frac{\gamma_D}{1-\gamma_D} U_O(R_b^o) \end{aligned}$$

iff

$$\begin{aligned} & (1-\gamma_D) \sum_{R \in supp(m_a)} m_a(R) U(R) + \gamma_D U_O(R_a^o) \\ \geq & (1-\gamma_D) \sum_{R \in supp(m_b)} m_b(R) U(R) + \gamma_D U_O(R_b^o). \square \end{aligned}$$

Proof of Proposition 4:

Since the comparison between any two sets, R and R' depends only on their best and worst elements, since the ordering of the best and the worst elements is the same as that of the singletons and coincides on the set $\boldsymbol{\mathcal{R}}_D$ and since the best and the worst singletons on both $\boldsymbol{\mathcal{R}}_D$ and $\Re_D$ are given by $\{\delta_{\bar{r}}\}$ and $\{\delta_{\underline{r}}\}$, we have that for any $(m, R^o) \in \mathbf{A}_D^o$,

$$(m(\{\delta_{\bar{r}}\}) = 1, \{\delta_{\bar{r}}\}) \succsim (m, R^o) \succsim (m(\{\delta_{\underline{r}}\}) = 1, \{\delta_{\underline{r}}\}),$$

or in the notation of Lemma 3,

$$(\bar{m}, \{\bar{\rho}\}) \succsim (m, R^o) \succsim \left(\underline{m}, \{\underline{\rho}\}\right)$$

By Corollary 1, we further have:

$$(\bar{m}, R^o) \succsim (m, R^o) \succsim (\underline{m}, R^o) \tag{23}$$

while by the non-triviality condition in Axiom 1, we have that there is an (m, R^o) for which either:

$$\begin{aligned} (m, R^o) &\succ (\underline{m}, R^o) \text{ or} \\ (\bar{m}, R^o) &\succ (m, R^o). \end{aligned}$$

If $(m, R^o) \succ (\underline{m}, R^o)$, we have by Corollary 1 that

$$(\bar{m}, \{\bar{\rho}\}) \succ (\underline{m}, R^o) \succsim \left(\underline{m}, \{\underline{\rho}\}\right)$$

and thus, $(\bar{m}, \{\bar{\rho}\}) \succ \left(\underline{m}, \{\underline{\rho}\}\right)$, whereas if $(\bar{m}, R^o) \succ (m, R^o)$, we have

$$(\bar{m}, \{\bar{\rho}\}) \succsim (\bar{m}, R^o) \succ \left(\underline{m}, \{\underline{\rho}\}\right)$$

and thus, again $(\bar{m}, \{\bar{\rho}\}) \succ \left(\underline{m}, \{\underline{\rho}\}\right)$. It follows that the γ_D identified in Lemma 3 is unique.

Next, observe that

$$(\bar{m}, \{\bar{\rho}\}) \sim (\underline{m}, \{\bar{\rho}\})$$

would contradict the non-triviality assumption imposed by Axiom 1. Indeed, it would imply, by Corollary 1 and by equation (23),

$$(\bar{m}, R^o) \sim (m, R^o) \sim (\underline{m}, R^o)$$

for all $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$ and all $R^o \in \Re_D$, in contradiction to Axiom 1. It follows that $(\bar{m}, \{\bar{\rho}\}) \succ (\underline{m}, \{\bar{\rho}\})$.

Next, assume that there is an $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$ and R and $R' \in \Re_D$ such that $(m, R) \succ (m, R')$. Thus,

$$(\underline{m}, \{\bar{\rho}\}) \succ \left(\underline{m}, \{\underline{\rho}\}\right),$$

which by Lemma 3 implies that $\gamma_D \in (0, 1)$ and, thus, the representation of Proposition 3.

Thus, we can set $U(R) = w^d\left(\underline{\rho}_R, \overline{\rho}_R\right)$ and $U_O(R) = w^o\left(\underline{\rho}_R, \overline{\rho}_R\right)$. By Proposition 3, these functions coincide on singleton sets $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$, i.e.,

$$w^d(\rho, \rho) = w^o(\rho, \rho)$$

and by Corollary 3, we can thus set $w^o(\rho, \rho) = \sum_{r \in R_D} u(r)\,\rho(r)$, where u is the utility function over outcomes identified in the Corollary. The uniqueness of the functions w^o and w^d follows from the respective uniqueness of U, U_O and u.

Hence the result of the Proposition obtains, provided that there is an $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$ and R and $R' \in \Re_D$ such that $(m, R) \succ (m, R')$ holds.

To complete the proof thus, suppose in a manner of contradiction that there are no $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$, R and $R' \in \Re_D$ such that $(m, R) \succ (m, R')$. We then have

$$(\underline{m}, \{\bar{\rho}\}) \sim (\underline{m}, \{\underline{\rho}\}),$$

resulting in $\gamma_D = 0$. Furthermore, $U_o(\{\delta_{\bar{r}}\}) = U_o(\{\delta_{\underline{r}}\})$. Axiom 5 then implies, $U_o(R^o) = U_o(\{\delta_{\underline{r}}\})$ for every $R^o \in \Re_D$. But by Axiom 4, we then obtain that for any $\lambda \in [0,1]$, any $\{\rho\} \in \boldsymbol{\mathcal{R}}_D$, and m and m' such that $m(\{\rho_1\}) = \lambda$, $m(\{\rho_2\}) = 1 - \lambda$, $m'(\{\rho\}) = 1$,

$$(\bar{m}, R^o) \sim (\underline{m}, R^o).$$

for some and thus, by Corollary 1, for any $R^o \in \Re_D$. At the same time, Axiom 5 gives us:

$$(\bar{m}, R^o) \sim (m, R^o) \sim (\underline{m}, R^o)$$

for any $m \in \Delta(\boldsymbol{\mathcal{R}}_D)$ for any $R^o \in \Re_D$ in contradiction to the non-triviality assumption in Axiom 1. □

8 Appendix: Axiomatizing Preferences over Algorithms with Learning

We here provide the necessary axioms which ensure that the behavior of the parameters of perception of unawareness and optimism satisfy the properties discussed in Section 5. To do so, we concentrate on the special case of classification presented in Section 2.2.1 In particular, cases are pairs of characteristics and class, (x, k), whereas actions are algorithms. For simplicity of exposition, we maintain the notation a for an action described as a mass function as in the axiomatization in Section 3.

Axiom 8 For any two data sets D and $D' \in \mathbb{D}$ and any two ρ_a, $\rho_b \in \Delta_{[0,1]}$,

$$(m_a(\{\rho_a\}) = 1, \{\rho_a\}) \succsim_D (m_b(\{\rho_b\}) = 1, \{\rho_b\})$$

iff

$$(m_a(\{\rho_a\}), \{\rho_a\}) \succsim_{D'} (m_b(\{\rho_b\}) = 1, \{\rho_b\}).$$

Axiom 8 implies that the utility function over outcomes is independent on the data set.

Let $\mathbb{D}(D_0)$ be the set of data sets which have the same set characteristics X_{D_0} as D_0.

Axiom 9a Let D and $D' \in \mathbb{D}(D_0)$ be such that $D' = D \circ \tilde{D}$. For any ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$ and any R^o,

$$(m_a(\{\rho,\rho'\}) = 1, R^o) \sim_D (m_b(\{\rho\}) = \alpha,\, m_b(\{\rho'\}) = 1-\alpha, R^o)$$

for some $\alpha \in [0,1]$ holds iff

$$(m_a(\{\rho,\rho'\}) = 1, R^o) \sim_{D'} (m_b(\{\rho\}) = \alpha,\, m_b(\{\rho'\}) = 1-\alpha, R^o)$$

Axiom 9a implies that the coefficients of optimism, α_D, for observed characteristics remain unchanged across data-sets: $\alpha_D = \alpha_{D'}$. Note that contrary to Axiom 8, Axiom 9 is conditional on the fact that the sets of characteristics are the same across the two data-sets. Thus, no surprises in the sense defined in the following section are observed. It is straightforward to formulate a similar axiom in regards to the degree of optimism relative to unobserved characteristics, α^o. Since, the cases in D_0 in which all characteristics X_D are observed are common to D and D', the set of surprises observed in D and in D' is identical, implying that $\alpha^o_D = \alpha^o_{D'}$:

Axiom 9b Let D and $D' \in \mathbb{D}(D_0)$ be such that $D' = D \circ \tilde{D}$. For any ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$ and any m,

$$(m, \{\rho,\rho'\}) \sim_D (m, \{\alpha\rho + (1-\alpha)\rho'\})$$

for some $\alpha \in [0,1]$ holds iff

$$(m, \{\rho,\rho'\}) \sim_{D'} (m, \{\alpha\rho + (1-\alpha)\rho'\})$$

Our final axiom guarantees that perceived unawareness decreases with the number of observations:

Axiom 10 For D and $D' \in \mathbb{D}(D_0)$ let the number of observations satisfy $N' > N$. For any ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$, if for some $\lambda \in [0,1]$,

$$(m_a(\{\rho\}) = 1, \{\rho'\}) \sim_D ((m_{a'}(\{\rho\}) = \lambda,\, m_{a'}(\{\rho'\}) = 1-\lambda), \{\lambda\rho + (1-\lambda)\rho'\})$$

then

$$(m_a(\{\rho\}) = 1, \{\rho'\}) \succ_{D'} ((m_{a'}(\{\rho\}) = \lambda,\, m_{a'}(\{\rho'\}) = 1-\lambda), \{\lambda\rho + (1-\lambda)\rho'\}).$$

Axiom 10 compares for the two data sets D and D' the action a, which assigns the better outcome ρ to the observed characteristics and the worse one to the "other" characteristics to the action a', which mixes ρ and ρ' with a factor λ on both the observed and on the "other" characteristics. If the decision maker is indifferent between these actions given D, this implies that the weight assigned to ρ' and thus to the "other" characteristics in the evaluation of a is identical to λ. The axiom then requires that this weight should decrease (with a becoming preferred to a') when the number of observations in the data set increases to that of D'.

Axiom 11 Consider $D \in \mathbb{D}_X$ such that $X_D \subset X$ and for $x^{new} \in X \backslash X_D$, $k \in K$, let $\tilde{D} = (x^{new}, k)$

$$D' = D \circ \tilde{D}$$

For any ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$, if for some $\lambda \in [0,1]$,

$$(m_a(\{\rho\}) = 1, \{\rho'\}) \sim_D ((m_{a'}(\{\rho\}) = \lambda,\, m_{a'}(\{\rho'\}) = 1-\lambda), \{\lambda\rho + (1-\lambda)\rho'\})$$

then

$$(m_a(\{\rho\}) = 1, \{\rho'\}) \prec_{D'} ((m_{a'}(\{\rho\}) = \lambda,\, m_{a'}(\{\rho'\}) = 1-\lambda), \{\lambda\rho + (1-\lambda)\rho'\}).$$

Axiom 11 is similar to Axiom 10 in that it compares for the two data-sets D and D' two actions: one in which mixing the good and the bad lottery ρ and ρ' in proportions λ and $(1-\lambda)$ occurs both on the set of observed, as well as for the "other" characteristics and a second one, in which the good lottery is obtained for sure on the set of known characteristics, but the bad one is realized on the unknown characteristics. In particular, λ is chosen so that it makes the decision maker indifferent between the two actions conditional on the information in data-set D. However, since differently from Axiom 10, in Axiom 11, data set D' differs from D by a surprise, i.e., by the observation of a new characteristic, the axiom stipulates that the decision maker will have a strict preference for the action which gives a constant mixture both on the set of observed, as well as for the "other" characteristics. This is tantamount to saying that the weight put on the "other" characteristics and thus, the perceived unawareness, will strictly increase in D' relative to D.

Our next Axiom formalizes the notion of a positive (negative) surprise and postulates the corresponding change in α^o:

Axiom 12 Consider $D \in \mathbb{D}_X$ such that $X_D \subset X$ and for $x^{new} \in X \backslash X_D$, $k \in K$, let $\tilde{D} = (x^{new}, k)$,

$$D' = D \circ \tilde{D}.$$

For ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$ and some m, let $\alpha \in [0,1]$ satisfy:

$$(m, \{\rho, \rho'\}) \sim_D (m, \{\alpha\rho + (1-\alpha)\rho'\})$$

(i) then

$$(m, \{\rho, \rho'\}) \succ_{D'} (m, \{\alpha\rho + (1-\alpha)\rho'\})$$

holds iff $\alpha \neq 1$ and $\tilde{D}$ is a positive surprise given D, i.e.,

$$\mathbf{a}^*(D) \succ_{D'|X_D} \mathbf{a}^*(D)(k, x_o) \text{ for all } \mathbf{a}^*(D) \in \mathbf{B}^*(D)$$

and

(ii)

$$(m, \{\rho, \rho'\}) \prec_{D'} (m, \{\alpha\rho + (1-\alpha)\rho'\})$$

holds iff $\alpha \neq 0$ and $\tilde{D}$ is a negative surprise given D, i.e.,

$$\mathbf{a}^*(D) \prec_{D'|X_D} \mathbf{a}^*(D)(k, x_o) \text{ for all } \mathbf{a}^*(D) \in \mathbf{B}^*(D)$$

Axiom 12 requires that if and only if a positive (negative) surprise has been experienced, the decision maker's evaluation of the outcome $\{\rho, \rho'\}$ on x_o increases (decreases) given D' as compared to D, i.e. the optimism parameter α^o is strictly higher (lower), (controlling for the boundaries 1 and 0) at D' than at D, $\alpha^o_{D'} > (<)\, \alpha^o_D$. It implies that if the surprise is neither positive, nor negative, or, if α^o_D is already on the respective boundary, the optimism parameter remains unchanged.

We next consider the case in which the data set is augmented by an observation of a so far unlabeled characteristic, $\hat{x}$, $\left(\hat{x}, \hat{k}\right)$.

Formally, define $X^U_D \subseteq X_D$ to be the set of unlabeled characteristics in the data set D:

$$X^U_D = \{x \in X_D \mid l_D(x) = \delta_{k^0}\}$$

These are the characteristics for which the recorded class is k^0. If $\hat{x} \in X^U_D$, then $f_D(\hat{x}) > 0$, but since $l_D(\hat{x}) = \delta_{k^0}$, in general, a set, $\mathbf{L}_D(\hat{x})$ is used as a prediction of the classification likelihood. For some $D \in \mathbb{D}(D_0)$, let thus $D' = D \circ \tilde{D}$.

It is straightforward to restate the definition of a positive / negative surprise by replacing x^{new} in definitions (10) and (11) by $\hat{x}$ with the same interpretation as above.

Behaviorally, given the axioms imposed on preferences, we can identify the type of surprise related to previously unlabeled characteristics by eliciting the decision maker's preferences between the optimal algorithm in D conditional on the set of observed labeled characteristics in D, $X_D \backslash X^U_D$ and the hypothetical act that assigns the prediction $a(\hat{x})$ and thus, the probability distribution over outcomes given by $a(\hat{x})\, l_{D'}(\hat{x}) = a\left(\hat{k}, \hat{x}\right)$ to any $X_D \backslash X^U_D$:

$$\mathbf{a}^*(D) \succ_{D'|X_D \backslash X^U_D} \left(\prec_{D'|X_D \backslash X^U_D}\right) \mathbf{a}^*(D)\left(\hat{k}, \hat{x}\right)$$

for all $a^*(D) \in \boldsymbol{B}^*(D)$.

Similarly to Axiom 12, Axiom 12a captures the fact that a positive (negative) surprise with respect to unlabeled characteristics might have an effect on the expectations the decision maker forms about his ability to predict the correct class for characteristics with unobserved classification. A positive (negative) surprise might make him more optimistic (pessimistic) increasing (decreasing) the degree of optimism α related to observed (but unlabeled) characteristics.

Axiom 12a Consider $D \in \mathbb{D}(D_0)$ and for $\hat{x} \in X^U_D$, let $\tilde{D} = \left(\hat{x}, \hat{k}\right)$ for some $\hat{k} \in K$ and

$$D' = D \circ \tilde{D}.$$

For ρ, $\rho' \in \Delta_{[0,1]}$ with $\rho \succ \rho'$ and some R^o, let $\alpha \in [0, 1]$ satisfy:

$$(m_a(\{\rho, \rho'\}) = 1, R^o) \sim_D (m_b(\{\rho\}) = \alpha,\, m_b(\{\rho'\}) = 1 - \alpha, R^o)$$

(i) then

$$(m_a(\{\rho, \rho'\}) = 1, R^o) \succ_{D'} (m_b(\{\rho\}) = \alpha,\, m_b(\{\rho'\}) = 1 - \alpha, R^o)$$

holds iff $\alpha \neq 1$ and $\tilde{D}$ is a positive surprise given D, i.e.,

$$\mathbf{a}^*(D) \succ_{D'|X_D \backslash \{\hat{x}\}} \mathbf{a}^*(D)\left(\hat{k}, \hat{x}\right) \text{ for all } \mathbf{a}^*(D) \in \mathbf{B}^*(D)$$

and

(ii)

$$(m_a(\{\rho, \rho'\}) = 1, R^o) \prec_{D'} (m_b(\{\rho\}) = \alpha, m_b(\{\rho'\}) = 1 - \alpha, R^o)$$

holds iff $\alpha \neq 0$ and $\tilde{D}$ is a negative surprise given D, i.e.,

$$\mathbf{a}^*(D) \prec_{D'|X_D} \mathbf{a}^*(D)(k, x_o) \text{ for all } \mathbf{a}^*(D) \in \mathbf{B}^*(D)$$

The following proposition combines Axioms 1-12a to represent preferences over the set of algorithms for the classification problem discussed in Section 2.2.1.

Proposition 5. *Consider a set of characteristics X and the associated class of data sets $\mathbb{D}_X$. Axioms 1–12a imply that there exist a von-Neumann-Morgenstern utility function u unique up to a positive-affine transformation, unique families of parameters of optimism $\{\alpha_D \in [0,1]\}_{D \in \mathbb{D}_X}$ relevant to the set of observed characteristics and $\{\alpha_D^o \in [0,1]\}_{D \in \mathbb{D}_X}$ relevant to the "other" characteristics as well as a family of parameters describing the perception of unawarenes for each data set $D \in \mathbb{D}(D_0)$, $\{\gamma_D \in (0;1)\}_{D \in \mathbb{D}_X}$ such that for each $D \in \mathbb{D}$, preferences over algorithms $\mathbf{a}(D)$, $\succsim_D$ are represented by (5), whereas the parameters of the representation satisfy:*

1. *for any $D_0 \in \mathbb{D}_X$ and D and $D' \in \mathbb{D}(D_0)$ with $X_D^U = X_{D'}^U$, $\alpha_D^o = \alpha_{D'}^o$, $\gamma_D > \gamma_{D'}$ iff $N' > N$ and $\alpha_D = \alpha_{D'}$;*

2. *for any $D_0 \in \mathbb{D}_X$, $D \in \mathbb{D}(D_0)$, $x \in X_D^U$, $k \in K$ and $D' = D \circ (x, k)$, $\alpha_D^o = \alpha_{D'}^o$, $\gamma_D > \gamma_{D'}$ and*

- *$\alpha_{D'} > \alpha_D$ iff (x, k) is a positive surprise given D and $\alpha_D^o \neq 1$*
- *$\alpha_{D'} < \alpha_D$ iff (x, k) is a negative surprise given D and $\alpha_D^o \neq 0$*

3. *for any $D_0 \in \mathbb{D}_X$, $D \in \mathbb{D}(D_0)$, $x \in X \backslash X_D$, $k \in K$ and $D' = D \circ (x, k)$, $\gamma_{D'} > \gamma_D$ and*

- *$\alpha_{D'}^o > \alpha_D^o$ iff (x, k) is a positive surprise given D and $\alpha_D^o \neq 1$*
- *$\alpha_{D'}^o < \alpha_D^o$ iff (x, k) is a negative surprise given D and $\alpha_D^o \neq 0$.*

References

Billot, A., Gilboa, I., Samet, D. and Schmeidler, D. (2005). Probabilities as Similarity-Weighted Frequencies. *Econometrica*, **73**, 1125–1136.

Chung, Y., Haas, P. J., Upfal, E. and Kraska, T. (2018). Unknown examples & machine learning model generalization. *ArXiv*, **abs/1808.08294**.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, **210** (C), 2–21.

Dominiak, A. and Guerdjikova, A. (2021). Pessimism and optimism towards new discoveries. *Theory and Decision*, **90**, 321–370.

Eichberger, J. and Guerdjikova, A. (2010). Case-based belief formation under ambiguity. *Mathematical Social Sciences*, **60**, 161–177.

— and — (2013). Ambiguity, Data and Preferences for Information - A Case-Based Approach. *Journal of Economic Theory*, **148**, 1433–1462.

— and Pasichnichenko, I. (2021). Decision-making with partial information. *Journal of Economic Theory*, **198**, 105369.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, **75**, 643–669.

Gilboa, I., Minardi, S., Samuelson, L. and Schmeidler, D. (2020). States and contingencies: How to understand savage without anyone being hanged. *Revue économique*, **71** (2), 365–386.

— and Schmeidler, D. (1989). Maxmin Expected Utility with a Non-Unique Prior. *Journal of Mathematical Economics*, **18**, 141–153.

— and — (2001). Reaction to Price Changes and Aspiration Level Adjustments. *Review of Economic Design*, **6**, 215–223.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning.* MIT Press.

Grabisch, M. (2016). *Set Functions, Games and Capacities in Decision Making*, *Theory and Decision Library C*, vol. 46. Springer, game theory, social choice, decsision theory, and optimization edn.

Grant, S., Guerdjikova, A. and Quiggin, J. (2020). Ambiguity and awareness: A coherent multiple prior model. *The B.E. Journal of Theoretical Economics:.*

—, Meneghel, I. and Tourky, R. (2017). Learning under unawareness. *Working Paper.*

— and Quiggin, J. (2006). Learning and discovery. *Risk and Sustainable Management Group Working Papers 151174*.

— and — (2013a). Bounded awareness, heuristics and the precautionary principle. *Journal of Economic Behavior & Organization*, **93**, 17–31.

— and — (2013b). Inductive reasoning about unawareness. *Economic Theory*, **54** (3), 717–755.

Jaffray, J.-Y. (1989). Linear Utility Theory for Belief Functions. *Operations Research Letters*, **8**, 107–112.

Karni, E. (2022). A theory-based decision model. *Journal of Economic Theory*, **201**.

— and Vierø, M.-L. (2013). Reverse Bayesianism: A choice-based theory of growing awareness. *American Economic Review*, **103**, 2790–2810.

— and Vierø, M.-L. (2017). Awareness of Unawareness: A Theory of Decision Making in the Face of Ignorance. *Journal of Economic Theory*, **168**, 301–328.

Kreps, D. M. (1988). Underground Classics in Economics, Bolder: Westview Press.

Savage, L. J. (1954). *Foundations of Statistics*. New York: Wiley.

Schipper, B. C. (2022). *Predicting the Unpredictable under Subjective Expected Utility*. Tech. rep., UC Davis.

Schmeidler, D. (1989). Subjective Probability and Expected Utility without Additivity. *Econometrica*, **57**, 571–587.

Short, W. M. (2018). The spacial metaphorics of ambiuity in roman culture. In M. Fontaine, C. McNamara and W. M. Short (eds.), *Quasi Labor Intus: Ambiguity in Latin Literature*, Leiden: Brill, p. 1–25.

Vierø, M.-L. (2021). An intertemporal model of growing awareness. *Journal of Economic Theory*, **197**, 105351.

von Neumann, J. and Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*. New Jersey: Princeton University Press.