Categorical Decision Theory
Extended abstract

Marcus Pivato [*]

March 29, 2024

Motivation. Individuals and societies must often make difficult decisions, which are fraught with uncertainty. How should an agent decide when faced with such uncertainty? This is the subject of a branch of theoretical economics called *decision theory.*

Bernoulli (1738) claimed that we should choose the alternative which yields the highest *expected utility.* But what justifies this claim? Savage (1954) showed that, if our decision-making process satisfies certain axioms (encoding basic properties of "consistency" and "rationality"), then it must maximize expected utility.

Savage posited a set $\mathcal{S}$ of possible "states of nature" and a set $\mathcal{X}$ of possible "outcomes". Each alternative is a function (an "act") mapping $\mathcal{S}$ to $\mathcal{X}$. He imagines an agent who must form preferences ($\succcurlyeq$) over these acts, even though she does not yet know the true state of nature. (The true state will be revealed only after she chooses some act.) Savage's theorem constructs a probability measure ρ on $\mathcal{S}$ and a utility function u on $\mathcal{X}$ such that, for any acts α and β, we have

$$\alpha \succcurlyeq \beta \quad \Longleftrightarrow \quad \int_{\mathcal{S}} u \circ \alpha \ \mathrm{d}\rho \geqslant \int_{\mathcal{S}} u \circ \beta \ \mathrm{d}\rho.$$

In other words: the agent prefers α to β if and only if the expected utility of α is at least as big as that of β. This is called a *subjective expected utility representation* of the agent's preferences, because the probability measure ρ is interpreted as representing her *subjective* beliefs. (There is no claim that ρ is in some sense the "objectively correct" probability measure on $\mathcal{S}$, or even that there *is* an "objectively correct" probability measure on $\mathcal{S}$.)

Savage's Theorem is considered the foundational result of modern decision theory. However, his approach raises several issues.

1. Savage assumed that $\mathcal{S}$ and $\mathcal{X}$ were arbitrary sets, and acts were arbitrary functions. But what if and $\mathcal{S}$ and $\mathcal{X}$ have some other mathematical structure (e.g. a sigma-algebra or a topology) and acts must preserve this structure? It is possible to prove "versions" of Savage's theorem specific to each mathematical environment. But it

[*] Centre d'Economie de la Sorbonne, Université Paris 1 Panthéon-Sorbonne

would be more efficient and intellectually satisfying to obtain a single theory which works in every environment.

2. In many applications, it is unrealistic to suppose that the agent can specify all possible states of nature or all possible outcomes "in advance". Thus, there is growing interest in developing decision theory *without* an explicit specification of $\mathcal{S}$ or $\mathcal{X}$. (This is particularly important in a branch of decision theory which deals with "unawareness".)

3. At different times, the same agent might be faced with many different sources of uncertainty (i.e. different instances of $\mathcal{S}$) and many different menus of outcomes (different instances of $\mathcal{X}$), in different combinations. We could, of course, use Savage's theorem to construct separate subjective expected utility representations for the agent's preferences for each possible combination of some $\mathcal{S}$ with some $\mathcal{X}$. But aside from the fact that this would be inelegant and redundant, there is no guarantee that the utility functions or probabilistic beliefs that we thus obtained would be consistent across the different possible combinations of $\mathcal{S}$ and $\mathcal{X}$. We would like a way to obtain a single *holistic* description of the agent's decisions over all of these possible decision problems, simultaneously.

In this talk, I will describe an ongoing research project in which I have reformulated decision theory using the tools of category theory, and derived a version of Savage's theorem which addresses all three of these issues.

Notation. For any category $\boldsymbol{\mathcal{C}}$, let $\boldsymbol{\mathcal{C}}^\circ$ denote the set of objects in $\boldsymbol{\mathcal{C}}$. For any objects $\mathcal{A}, \mathcal{B} \in \boldsymbol{\mathcal{C}}^\circ$, let $\overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{A}, \mathcal{B})$ denote the set of $\boldsymbol{\mathcal{C}}$-morphisms from $\mathcal{A}$ to $\mathcal{B}$.

Decision Environments. Let $\boldsymbol{\mathcal{C}}$ be a category. A *decision environment* on $\boldsymbol{\mathcal{C}}$ is an ordered pair $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$, where $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{X}}$ are subcategories of $\boldsymbol{\mathcal{C}}$.[1] We shall interpret the objects of $\boldsymbol{\mathcal{S}}^\circ$ as "abstract state spaces", and interpret objects of $\boldsymbol{\mathcal{X}}^\circ$ as "abstract outcome spaces". However, if $\boldsymbol{\mathcal{C}}$ is an abstract category, then these might not actually be spaces of any kind. For this reason, we shall refer to the objects of $\boldsymbol{\mathcal{S}}^\circ$ as *state places* and the objects of $\boldsymbol{\mathcal{X}}^\circ$ as *outcome places*.

For any state place $\mathcal{S}$ in $\boldsymbol{\mathcal{S}}^\circ$ and outcome place $\mathcal{X}$ in $\boldsymbol{\mathcal{X}}^\circ$, the morphisms in $\overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$ represent "abstract acts" —these are devices that somehow transform the abstract "states" in $\mathcal{S}$ into abstract "outcomes" in $\mathcal{X}$. For simplicity, we shall call them *acts*. If $\boldsymbol{\mathcal{C}}$ was a concrete category, then $\mathcal{S}$ and $\mathcal{X}$ would be sets, and the acts in $\overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$ would be functions. But we shall not assume this.

Heuristically, each state place in $\boldsymbol{\mathcal{S}}^\circ$ represents an uncertain situation. Suppose that the agent has "beliefs" about these uncertain situations. For any state places $\mathcal{S}_1, \mathcal{S}_2 \in \boldsymbol{\mathcal{S}}^\circ$, we interpret each element of $\overrightarrow{\boldsymbol{\mathcal{S}}}(\mathcal{S}_1, \mathcal{S}_2)$ as a $\boldsymbol{\mathcal{C}}$-morphism from $\mathcal{S}_1$ to $\mathcal{S}_2$ that is somehow "compatible" with her beliefs about $\mathcal{S}_1$ and $\mathcal{S}_2$. (We shall make this precise shortly.)

[1] That is: $\boldsymbol{\mathcal{S}}^\circ \subseteq \boldsymbol{\mathcal{C}}^\circ$ and $\boldsymbol{\mathcal{X}}^\circ \subseteq \boldsymbol{\mathcal{C}}^\circ$. Furthermore, for any $\mathcal{S}_1, \mathcal{S}_2 \in \boldsymbol{\mathcal{S}}^\circ$, we have $\overrightarrow{\boldsymbol{\mathcal{S}}}(\mathcal{S}_1, \mathcal{S}_2) \subseteq \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}_1, \mathcal{S}_2)$; likewise, for any $\mathcal{X}_1, \mathcal{X}_2 \in \boldsymbol{\mathcal{X}}^\circ$, we have $\overrightarrow{\boldsymbol{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2) \subseteq \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{X}_1, \mathcal{X}_2)$.

For example, if $\mathcal{S}_1$ and $\mathcal{S}_2$ were measurable spaces, and the agent's beliefs took the form of probability measures, then $\overrightarrow{\boldsymbol{\mathcal{S}}}(\mathcal{S}_1, \mathcal{S}_2)$ could be the set of *measure-preserving* functions from $\mathcal{S}_1$ into $\mathcal{S}_2$. However, we shall not (yet) commit to any formal model of the agent's beliefs (e.g. as probabilities), so we shall not (yet) impose any restrictions on the sort of morphisms that can appear in $\overrightarrow{\boldsymbol{\mathcal{S}}}(\mathcal{S}_1, \mathcal{S}_2)$.

Meanwhile, each outcome place in $\boldsymbol{\mathcal{X}}^\circ$ represents an abstract "menu" of possible outcomes. Suppose that the agent has "tastes" over these menus. For any $\mathcal{X}_1, \mathcal{X}_2 \in \boldsymbol{\mathcal{X}}^\circ$, we interpret each element of $\overrightarrow{\boldsymbol{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2)$ as a $\boldsymbol{\mathcal{C}}$-morphism from $\mathcal{X}_1$ to $\mathcal{X}_2$ that is somehow "compatible" with her tastes over $\mathcal{X}_1$ and $\mathcal{X}_2$. (We shall make this precise shortly.) For example, if her tastes took the form of preference orders on $\mathcal{X}_1$ and $\mathcal{X}_2$, then $\overrightarrow{\boldsymbol{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2)$ could be the set of *order-preserving* $\boldsymbol{\mathcal{C}}$-morphisms from $\mathcal{X}_1$ into $\mathcal{X}_2$. However, we shall not (yet) commit to any formal model of the agent's tastes (e.g. in terms of preference orders or utility functions), so we shall not (yet) impose any restrictions on the sort of morphisms that can appear in $\overrightarrow{\boldsymbol{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2)$.

Ex ante preference structures. Let $\boldsymbol{\mathcal{C}}$ be a category, and let $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$ be a decision environment on $\boldsymbol{\mathcal{C}}$. For every $\mathcal{S} \in \boldsymbol{\mathcal{S}}^\circ$ and $\mathcal{X} \in \boldsymbol{\mathcal{X}}^\circ$, let $\succeq_{\mathcal{X}}^{\mathcal{S}}$ be a weak order[2] on $\overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$. The collection $\succeq^{\text{xa}} = \{\succeq_{\mathcal{X}}^{\mathcal{S}};\ \mathcal{S} \in \boldsymbol{\mathcal{S}}^\circ \text{ and } \mathcal{X} \in \boldsymbol{\mathcal{X}}^\circ\}$ is an *ex ante preference structure* on $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$ if it is "compatible" with the morphisms of $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{X}}$ in the following two senses.

(BP) For any state places $\mathcal{S}_1, \mathcal{S}_2 \in \boldsymbol{\mathcal{S}}^\circ$, any morphism $\phi \in \overrightarrow{\boldsymbol{\mathcal{S}}}(\mathcal{S}_1, \mathcal{S}_2)$, any outcome place $\mathcal{X} \in \boldsymbol{\mathcal{X}}^\circ$, and any acts $\alpha, \beta \in \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}_2, \mathcal{X})$, we have

$$\left(\alpha \succeq_{\mathcal{X}}^{\mathcal{S}_2} \beta\right) \iff \left(\alpha \circ \phi \succeq_{\mathcal{X}}^{\mathcal{S}_1} \beta \circ \phi\right).$$

(TP) For any outcome places $\mathcal{X}_1, \mathcal{X}_2 \in \boldsymbol{\mathcal{X}}^\circ$, any morphism $\phi \in \overrightarrow{\boldsymbol{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2)$, any state place $\mathcal{S} \in \boldsymbol{\mathcal{S}}^\circ$, and any acts $\alpha, \beta \in \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X}_1)$, we have

$$\left(\alpha \succeq_{\mathcal{X}_1}^{\mathcal{S}} \beta\right) \iff \left(\phi \circ \alpha \succeq_{\mathcal{X}_2}^{\mathcal{S}} \phi \circ \beta\right).$$

Compatibility condition (BP) formalizes our earlier informal assertion that the morphisms of the subcategory $\boldsymbol{\mathcal{S}}$ are "belief-preserving". Likewise, condition (TP) formalizes our earlier assertion that the morphisms of the subcategory $\boldsymbol{\mathcal{X}}$ are "taste-preserving". Our main research question is as follows:

> *Under what conditions does an ex ante preference structure admit a subjective expected utility representation?*

Background conditions on $\boldsymbol{\mathcal{C}}$. Assume that $\boldsymbol{\mathcal{C}}$ has an initial object ($\mathbf{0}_{\boldsymbol{\mathcal{C}}}$) and a terminal object ($\mathbf{1}_{\boldsymbol{\mathcal{C}}}$). For any object $\mathcal{X} \in \boldsymbol{\mathcal{C}}^\circ$, let $\underline{\mathcal{X}}$ denote the set of global elements of $\mathcal{X}$ (i.e. morphisms from $\mathbf{1}_{\boldsymbol{\mathcal{C}}}$ into $\mathcal{X}$). We will denote generic elements of $\underline{\mathcal{X}}$ by $\underline{x}$, $\underline{y}$, etc. Let

[2] That is: a complete, transitive, reflexive binary relation.

us say that $\mathcal{X}$ is *inhabited* if $\underline{\mathcal{X}} \neq \varnothing$. Say that the category $\boldsymbol{\mathcal{C}}$ is *well-inhabited* if (1) $\mathbf{0}_{\boldsymbol{\mathcal{C}}}$ is uninhabited; (2) every *non*-initial object is inhabited; and (3) There is some $\mathcal{X} \in \boldsymbol{\mathcal{C}}^{\circ}$ with $|\underline{\mathcal{X}}| \geqslant 2$. Say that $\boldsymbol{\mathcal{C}}$ is *hospitable* if it is well-inhabited, pullback-complete, and has pullback-stable coproducts.

Structural conditions. We need the decision environment $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$ to satisfy three conditions.

(S1) For any two distinct state places $\mathcal{S}_1$ and $\mathcal{S}_2$ in $\boldsymbol{\mathcal{S}}^{\circ}$, there is a third state place $\mathcal{S}$ in $\boldsymbol{\mathcal{S}}^{\circ}$ and $\boldsymbol{\mathcal{S}}$-morphisms $\mathcal{S} \longrightarrow \mathcal{S}_1$ and $\mathcal{S} \longrightarrow \mathcal{S}_2$.

(S2) For any two distinct outcome places $\mathcal{X}_1$ and $\mathcal{X}_2$ in $\boldsymbol{\mathcal{X}}^{\circ}$, there is a third outcome place $\mathcal{X}$ in $\boldsymbol{\mathcal{X}}^{\circ}$ and $\boldsymbol{\mathcal{X}}$-morphisms $\mathcal{X}_1 \longrightarrow \mathcal{X}$ and $\mathcal{X}_2 \longrightarrow \mathcal{X}$.

(S3) $\boldsymbol{\mathcal{S}}$ is "closed under pullbacks" in the following sense. Suppose we have a pullback diagram in the category $\boldsymbol{\mathcal{C}}$:

$$\begin{tikzcd}
\mathcal{S}_* \arrow[r, dashed, "\phi_*"] \arrow[d, dashed, "\psi_*"'] \arrow[dr, phantom, "\lrcorner", very near start] & \mathcal{S}_1 \arrow[d, "\psi"] \\
\mathcal{S}_2 \arrow[r, "\phi"'] & \mathcal{S}_3
\end{tikzcd}$$

If $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ are all in $\boldsymbol{\mathcal{S}}^{\circ}$, and ϕ and ψ are $\boldsymbol{\mathcal{S}}$-morphisms, then $\mathcal{S}_*$ is *also* in $\boldsymbol{\mathcal{S}}^{\circ}$, and ϕ_* and ψ_* are also $\boldsymbol{\mathcal{S}}$-morphisms.

Quasiconstant morphisms. Let $\mathcal{S}, \mathcal{X} \in \boldsymbol{\mathcal{C}}^{\circ}$. Let us say a morphism $\kappa \in \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$ is *quasiconstant* if there is some $\underline{x} \in \underline{\mathcal{X}}$ such that the following diagram commutes:

$$\begin{tikzcd}
\mathcal{S} \arrow[d, "!_{\mathcal{S}}"'] \arrow[dr, "\kappa"] & \\
\mathbf{1}_{\boldsymbol{\mathcal{C}}} \arrow[r, dashed, "\underline{x}"'] & \mathcal{X}
\end{tikzcd}$$

Let us say that $\underline{x}$ is the *value* of κ. Let $\mathcal{K}(\mathcal{S}, \mathcal{X})$ be the set of all quasiconstant morphisms from $\mathcal{S}$ to $\mathcal{X}$.

Preference axioms. We need the ex ante preference structure $\overset{\mathrm{xa}}{\succeq}$ to satisfy four axioms. Here is the first one.

(A1) (*Ex post preferences*) For any $\mathcal{X} \in \boldsymbol{\mathcal{X}}^{\circ}$, there is a nontrivial preference order $\overset{\mathrm{xp}}{\succeq}_{\mathcal{X}}$ on $\underline{\mathcal{X}}$ such that, for any $\mathcal{S} \in \boldsymbol{\mathcal{S}}^{\circ}$ and any $\kappa_1, \kappa_2 \in \mathcal{K}(\mathcal{S}, \mathcal{X})$ with values $\underline{x}_1, \underline{x}_2$, we have $\kappa_1 \succeq^{\mathcal{S}}_{\mathcal{X}} \kappa_2$ if and only if $\underline{x}_1 \overset{\mathrm{xp}}{\succeq}_{\mathcal{X}} \underline{x}_2$.

Let $\alpha, \beta \in \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$. Let us say that α ***statewise dominates*** β, and write $\alpha \overset{\mathrm{dom}}{\succeq} \beta$, if $\alpha \circ \underline{s} \overset{\mathrm{xp}}{\succeq}_{\mathcal{X}} \alpha \circ \underline{s}$ for any $\underline{s} \in \underline{\mathcal{S}}$, where $\overset{\mathrm{xp}}{\succeq}_{\mathcal{X}}$ is the ex post preference relation on $\underline{\mathcal{X}}$ posited by axiom (A1). Here is the second axiom.

(A2) (*Statewise dominance*) For any $\mathcal{S} \in \boldsymbol{\mathcal{S}}^\circ$ and $\mathcal{X} \in \boldsymbol{\mathcal{X}}^\circ$, and any $\alpha, \beta \in \overrightarrow{\boldsymbol{\mathcal{C}}}(\mathcal{S}, \mathcal{X})$, if $\alpha \succeq^{\mathrm{dom}} \beta$, then $\alpha \succeq^{\mathcal{S}}_{\mathcal{X}} \beta$.

Due to space constraints, we can only state the last two axioms and the main result informally.

(A3) For any $\mathcal{X} \in \boldsymbol{\mathcal{X}}^\circ$, the preference relation $\succeq^{\mathrm{xp}}_{\mathcal{X}}$ is *Archimedean.* Roughly speaking: for any $\underline{w}, \underline{x}, \underline{y}, \underline{z} \in \underline{\mathcal{X}}$, the value difference between $\underline{w}$ and $\underline{x}$ cannot be "infinitesimal" relative to the value difference between $\underline{y}$ and $\underline{z}$.

(A4) Tradeoffs between outcomes are consistent: if the agent is indifferent between "trading outcome $\underline{w}$ for $\underline{x}$" and "trading outcome $\underline{y}$ for $\underline{z}$", but strictly prefers $\underline{w}'$ to $\underline{w}$, then she *cannot* be indifferent between "trading $\underline{w}'$ for $\underline{x}$" and "trading $\underline{y}$ for $\underline{z}$".

Theorem. (Informal statement) *Let $\boldsymbol{\mathcal{C}}$ be any hospitable category, let $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$ be a decision environment on $\boldsymbol{\mathcal{C}}$ satisfying structural conditions* (S1)-(S3)*, and let $\succeq^{\mathrm{xa}}$ be a solvable ex ante preference structure on $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}})$. Then $\succeq^{\mathrm{xa}}$ admits a "subjective expected utility representation" if and only if it satisfies axioms* (A1)-(A4)*. In this representation, the probabilistic "beliefs" over each state place are unique. The "utility function" on each outcome place represents the ex post preferences of the agent, and is unique up to positive affine transformations.*

In the informal statement of axioms (A3) and (A4) and the theorem, many terms appear in quotation marks, because they have not yet been formally defined. Indeed, although the meaning of these terms is fairly obvious in a concrete category like the category of sets and functions or the category of measurable spaces and measurable maps, it is not clear what these terms could even *mean* in an abstract category. So in order to formally state the theorem, or even to formulate axioms (A3) and (A4), we need a theoretical framework in which these terms can be precisely defined. Much of the work in this project involves finding a way to define these terms which makes sense in an abstract category.