Language-based Decisions

Adam Bjorndahl
Department of Philosophy
Carnegie Mellon University
Pittsburgh, USA
abjorn@cmu.edu

Joseph Y. Halpern
Department of Computer Science
Cornell Univeristy
Ithaca, USA
halpern@cs.cornell.edu

In Savage's classic decision-theoretic framework [12], actions are formally defined as functions from states to outcomes. But where do the state space and outcome space come from? Expanding on recent work by Blume, Easley, and Halpern [3], we consider a language-based framework in which actions are identified with (conditional) descriptions in a simple underlying language, while states and outcomes (along with probabilities and utilities) are constructed as part of a representation theorem. Our work expands the role of language from that in [3] by using it not only for the *conditions* that determine which actions are taken, but also the *effects*. More precisely, we take the set of actions to be built from those of the form $do(\varphi)$, for formulas φ in the underlying language. This presents a problem: how do we interpret the result of $do(\varphi)$ when φ is underspecified (i.e., compatible with multiple states)? We answer this using tools familiar from the semantics of counterfactuals [13]: roughly speaking, $do(\varphi)$ maps each state to the "closest" φ-state. This notion of "closest" is also something we construct as part of the representation theorem; in effect, then, we prove that (under appropriate assumptions) the agent is acting *as if* each underspecified action is first made definite and then evaluated (i.e., by maximizing expected utility). Of course, actions in the real world are often not presented in a fully precise manner, yet agents reason about and form preferences among them all the same. Our work brings the abstract tools of decision theory into closer contact with such real-world scenarios.

1 Motivation

In Savage's classic decision-theoretic framework [12] *actions* are formally defined as functions from *states* to *outcomes*. States are conceptualized as encoding the possible uncertainty the decision-maker may have about the world, while outcomes correspond intuitively to the payoff-relevant ways things might turn out. Thus, an action α can be viewed as a kind of long list: for each way the world might be (i.e., each state s), α specifies what will happen—namely, the outcome $\alpha(s)$—in case action α is actually performed in state s.

One might ask: where do the state space and outcome space come from? Is it reasonable to model an agent using a mathematical apparatus they presumably have no access to? Questions like these tap into a long tradition of challenging the idealizations involved in models like Savage's (see, e.g., [1, 2, 4, 5, 6, 7, 8, 9, 10, 14]). One response might be that we are not trying to *duplicate* the decision-making process going on "in the agent's head", but rather to *represent* it, mathematically—to show that under certain conditions it can be tracked with a certain type of formalism (in this case, as a form of expected utility maximization).

Although this reply might assuage some worries about the use of abstract mathematical frameworks for reasoning about decision making in general, it remains problematic that actions—the objects over which agents are supposed to "reveal" their preferences, through concrete, binary choices—cannot themselves be described except by reference to the background state and outcome spaces, which might not

Submitted to:
TARK 2021

be the states and outcomes that the agent is actually thinking of. In such models, although outcomes are what agents are supposed to ultimately care about, actions are the *means* by which they bring outcomes about. This makes an agent's preferences regarding actions arguably the closest point of contact that these models have to the empirical, observable reality of choosing between alternatives. Indeed, this interpretation of actions is what underlies many of the intuitions brought to bear to justify the various axioms of decision making that Savage postulates and relies upon to prove his celebrated representation theorem.

The concern with where the states and outcomes are coming from motivated Blume, Easley, and Halpern [3] (henceforth BEH) to consider a model where acts and language are taken to be primary in a sense that we explain shortly, while the state and outcome space are constructed as part of the representation rather than specified exogenously. In more detail, BEH assumed that acts were programs in a simple programming language formed by closing off a set of primitive programs using **if** ... **then** ... **else** ..., so that if a and b are programs and t is a test (intuitively, a formula in a propositional language), then **if** t **then** a **else** b is a program. Thus, rather than conditioning actions on events (i.e., subsets of a state space), they are conditioned on *descriptions* of events, namely, tests. This approach allows BEH to not only circumvent a fixed, exogenous specification of the state space and outcome space (instead, they are constructed as part of a representation theorem, and programs are identified with maps from from these states to outcomes), but also (as they illustrate with several examples) makes it possible to capture a variety of *framing* effects, which basically derive from a mismatch between how the modeler conceives of the world and how the agent does, as manifested in different ways that descriptions of events might map onto actual events.

Our work is perhaps best understood as an extension of their work in which the role of language is even more central. Specifically, while BEH allowed arbitrary primitive programs, we take the primitive programs to have the form $do(\varphi)$, where φ is a formula. The $do(\varphi)$ notation follows Pearl [11]; intuitively, $do(\varphi)$ means that the agent somehow makes φ true. Note that this action is somewhat underspecified; it does not say what else becomes true as a result of φ being true; for example, if ψ is independent of φ, it does not tell us whether ψ or $\neg\psi$ is true. In our representation theorem, we assume that the agent has a way of specifying the effects of $do(\varphi)$. In more detail, we take states in our state space to be characterized by formulas in the language (this is similar to the canonical model used in BEH's representation theorem), and take the outcome space to be the same as the state space, so that a program maps states to states. As part of the representation theorem, the agent must decide what state $do(\varphi)$ maps each state ω to. We follow standard approaches to giving semantics to counterfactuals [13] by taking $do(\varphi)$ to map ω to the "closest" state to ω (according to some measure of closeness) where φ is true. Of course, what counts as "closest" depends on the agent's subjective view of the world, and is constructed from their preferences over acts.

This approach allows us to model choices in a way that seems to us closer to how agents perceive and reason about the options available to them. To illustrate, consider a policy-maker trying to decide whether to raise the minimum wage to \$15 or to leave it as is. In our framework, this amounts to comparing the acts $do(MW = \$15)$ and $do(true)$ (where $do(true)$ amounts to doing nothing). Of course, different agents may disagree about the side-effects of increasing the minimum wage (businesses may close, there may be more automation so jobs may be lost, and so on). This amounts to saying that different agents will interpret $do(MW = \$15)$ differently as a function from states to states, although all will agree that it will result in a state where the minimum wage is \$15.[1] We can also express contingent policies in our

[1] We remark that in this paper we consider only the single-agent case, but we find the multi-agent case, and specifically the effect of disagreements about what the closest state is, an exciting direction for future work.

framework, for example, raising the minimum wage if the economy is healthy.

By making both the acts and the test conditions formulas, we can capture framing and coarseness effects not only in the test conditions, but also in the choices. For example, we might imagine agents reacting differently to statements like "we will require that every citizen is paid at least $15 dollars for each hour they work" versus "we will require every business owner to pay their employees at least $15 for each hour they work", even if we can see that these are equivalent statements. Our framework would allow this.

The rest of this paper is organized as follows. We present our approach as an extension of the work of BEH. This has the benefit of allowing us to apply their representation theorem directly and focus our efforts on the novel aspects of our extension. We begin in Section 2 by reviewing the relevant definitions from BEH and augmenting them with the new ones we need to capture language-based, underspecified effects of actions. Then in Section 3 we articulate the representation theorem we are aiming at, introduce decision-theoretic axioms that allow us to achieve it—including axioms from BEH (Section 3.1) as well as several new axioms (Section 3.2)—and finally prove the theorem (Section 3.3). Section 4 concludes with a discussion of future work. Appendix A collects proofs omitted from the main text.

2 Language, Actions, and Models

Our first step is to import the relevant definitions from BEH so as to present our extension of their work in context. In order to emphasize the changes that we make and to streamline the presentation, we alter some of their notation and terminology, and focus on the special case of their system without randomization.

Let Φ denote a finite set of *primitive propositions*, and $\mathcal{L} = \mathcal{L}(\Phi)$ the propositional language consisting of all Boolean combinations of these primitives. Although of course it is possible (and interesting) to consider other languages, in this work we focus on languages of this form as the *underlying language of action*—intuitively, the language in which both the conditions and the results of actions are specified.

A **basic model (over $\mathcal{L}(\Phi)$)** is a tuple $M = (\Omega, [\![\cdot]\!]_M)$ where Ω is a nonempty set of *states* and $[\![\cdot]\!]_M : \Phi \to 2^\Omega$ is a *valuation function*. The valuation is recursively extended to all formulas in $\mathcal{L}$ in the usual way. Intuitively, $[\![\varphi]\!]_M$ is the set of states where φ is true. Using $[\![\cdot]\!]_M$ allows us to interpret descriptions in the language $\mathcal{L}$ (what BEH call "tests") as events: φ is interpreted as the subset $[\![\varphi]\!]_M \subseteq \Omega$ of the state space Ω. We sometimes drop the subscript when the model is clear from context, and write $\omega \models \varphi$ for $\omega \in [\![\varphi]\!]$. We say that φ is *satisfiable in M* if $[\![\varphi]\!]_M \neq \emptyset$ and that φ is *valid in M* if $[\![\varphi]\!]_M = \Omega$, and write $\models \varphi$ to indicate that φ is valid in all basic models. Finally, we define the *theory of ω (in M)* to be the set of all formulas true at ω, denoted $Th(\omega) = \{\varphi : \omega \models \varphi\}$, and write $\omega \equiv \omega'$ iff $Th(\omega) = Th(\omega')$.

Up to now, everything we have defined has followed BEH exactly—their "primitive tests" are our primitive propositions Φ; their "tests" are our formulas $\mathcal{L}(\Phi)$; their "test interpretations" are our valuations $[\![\cdot]\!]_M$. Next we define our version of their "primitive choices". This is where our development begins to diverge, since we take these to be actions of the form $do(\varphi)$; in other words, we specify primitive choices using the same underlying language $\mathcal{L}(\Phi)$ that corresponds to tests, rather than treating them as a brand new set of primitives.

Formally, given a finite set of formulas $F \subseteq \mathcal{L}$, the set of **actions (over F)**, denoted by $\mathcal{A}_F$, is defined recursively as follows: for each $\varphi \in F$, $do(\varphi)$ is an action (called a *primitive action*), and for all $\psi \in \mathcal{L}$ and $\alpha, \beta \in \mathcal{A}_\mathcal{L}$, **if ψ then α else β** is an action. Following BEH, we take F to be finite (who take the set of primitive choices to be finite). It is also convenient because it allows us to exclude logical inconsistencies from F, obviating the need to interpret actions like $do(false)$. For the propositional languages under

consideration in this paper, up to logical equivalence, there are only finitely many formulas in any case.

Naturally, we also wish to *interpret* our actions in a way that respects their connection to the underlying language. This is the topic we turn to next.

2.1 Selection models

In a given basic model M, we want $do(\varphi)$ to correspond to a function whose range is contained in $[\![\varphi]\!]_M$, the set of φ-states. Thus, we restrict our attention to basic models in which each $\varphi \in F$ is satisfiable—in this case we say that M is **F-rich**. But this is not enough: as discussed, $do(\varphi)$ is underspecified; it does not in general determine a unique function. In order to interpret such actions and compare them to others, we must in some sense "fill in" the missing details. We formalize this with the concept of a **selection model (for F)**, which is a basic model $M = (\Omega, [\![\cdot]\!]_M)$ together with a *selection function (for M)* $c : \Omega \times F \to \Omega$ satisfying $c(\omega, \varphi) \in [\![\varphi]\!]_M$.

Selection functions were introduced by Stalnaker [13] as a mechanism to interpret counterfactual conditionals. Following this tradition, we think of $c(\omega, \varphi)$ as representing the "closest" state to ω where φ is true. There are many other properties one might insist c have, aside from $c(\omega, \varphi) \in [\![\varphi]\!]$ (which is called **success**). For example, one may require that if $\omega \in [\![\varphi]\!]$, then $c(\omega, \varphi) = \omega$ (i.e., if φ is true in ω, then the closest state to ω where φ is true is ω itself); this property is called **centering**.

In this paper we will also consider a relatively strong condition on c, namely, that it is derived from a parametrized family of *well-orders*[2] on the state space, one for each state: $\leq := \{\leq_\omega : \omega \in \Omega\}$. Intuitively, $\omega_1 \leq_\omega \omega_2$ says "ω_1 is at least as close to ω as ω_2 is". We say that a selection function c is **induced by** $\leq$ if $c(\omega, \varphi)$ always outputs the $\leq_\omega$-minimal element of $[\![\varphi]\!]$. We call $\leq$ **centered** if, for each $\omega \in \Omega$, the $\leq_\omega$-minimal element of Ω is ω (in which case it is also easy to see that the induced selection function satisfies centering). Finally, we say that $\leq$ is **language-based** if the relations $\leqq_\omega$ on the quotient $\Omega/\equiv$ given by

$$[\omega_1] \leqq_\omega [\omega_2] \text{ iff } \omega_1 \leq_\omega \omega_2$$

Can we drop this moreover/

are well-defined well-orders, and [redacted]. Note that in this case $\omega \equiv \omega'$ implies $c(\omega, \varphi) \equiv c(\omega', \varphi)$.[3] Intuitively, if $\leq$ is language-based then what counts as the closest state essentially depends only on the formulas that are true at a state. We cannot have two states ω_1 and ω_2 that agree on all formulas (so that $\omega_1 \equiv \omega_2$) and a third state ω_3 that does not agree with ω_1 and ω_2 on all formulas such that ω_3 is between ω_1 and ω_2 in terms of distance from some state ω (i.e., we cannot have $\omega_1 \leq_\omega \omega_3 \leq_\omega \omega_2$).

The purpose of the selection function in our models is to take an underspecified transition from states to states and "resolve the ambiguity". Specifically, given a transition that starts in state ω and ends up in a φ-state, the selection function c can then by applied to specify the exact φ-state, namely $c(\omega, \varphi)$, where it actually ends up. In this way, given a basic, F-rich model M, each action of the form $do(\varphi)$ can be interpreted in any selection model (M, c) based on M as a function $[\![do(\varphi)]\!]_{M,c} : \Omega \to \Omega$ defined by:

$$[\![do(\varphi)]\!]_{M,c}(\omega) = c(\omega, \varphi).$$

[2] A binary relation $\leq$ on a set is called a *linear order* if it is complete, transitive, and antisymmetric (i.e., $x \leq y$ and $y \leq x$ implies $x = y$). A *well-order* is a linear order in which every nonempty subset has a least element.

[3] Here's why: since $c(\omega, \varphi)$ is the $\leq_\omega$-minimal element of $[\![\varphi]\!]$, it must also be that $[c(\omega, \varphi)]$ is the $\leqq_\omega$-minimal element of $\{[\omega''] : \omega'' \models \varphi\}$. Similarly, $[c(\omega', \varphi)]$ is the $\leqq_{\omega'}$-minimal element of $\{[\omega''] : \omega'' \models \varphi\}$. Since $\leqq_\omega = \leqq_{\omega'}$, these must coincide, so we have $[c(\omega, \varphi)] = [c(\omega', \varphi)]$.

Of course, we can extend this interpretation to all actions in $\mathcal{A}_F$ in the obvious way (and exactly as BEH do):

$$[\![\textbf{if } \psi \textbf{ then } \alpha \textbf{ else } \beta]\!]_{M,c}(\omega) = \begin{cases} [\![\alpha]\!]_{M,c}(\omega) & \text{if } \omega \in [\![\psi]\!] \\ [\![\beta]\!]_{M,c}(\omega) & \text{if } \omega \notin [\![\psi]\!]. \end{cases}$$

3 Representation

We begin as usual with a binary relation $\succeq$ on $\mathcal{A}_F$, where $\alpha \succeq \beta$ says that α is "at least as good as" β. Following standard conventions, we define $\alpha \succ \beta$ as an abbreviation for $\alpha \succeq \beta$ and $\beta \not\succeq \alpha$, and $\alpha \sim \beta$ for $\alpha \succeq \beta$ and $\beta \succeq \alpha$, representing "strict preference" and "indifference", respectively. We also assume that $\succeq$ is *complete*, that is, all elements are comparable, so that for all acts α and β, either $\alpha \succeq \beta$ or $\beta \succeq \alpha$. Although BEH consider incomplete relations, we focus here on the simpler case of complete relations in order to streamline the presentation and highlight the novel components of our model.

A **language-based SEU (Subjective Expected Utility) representation** for a relation $\succeq$ on $\mathcal{A}_F$ is a finite selection model (M,c) together with a probability measure π on Ω and a *utility function* $u : \Omega \to \mathbb{R}$ such that, for all $\alpha, \beta \in \mathcal{A}_F$,

$$\alpha \succeq \beta \Leftrightarrow \sum_{\omega \in \Omega} \pi(\omega) \cdot u([\![\alpha]\!]_{M,c}(\omega)) \geq \sum_{\omega \in \Omega} \pi(\omega) \cdot u([\![\beta]\!]_{M,c}(\omega)). \tag{1}$$

We note the key differences between the representation theorem BEH establish and what we are aiming at. First, their result produces a separate outcome space and state space, whereas for us, these spaces coincide. More importantly, their result treats "primitive choices" (namely, our actions $do(\varphi)$, for $\varphi \in F$) as true primitives in the sense that each is assigned to an *arbitrary* function from states to outcomes. By contrast, we want to respect the structure of an action like $do(\varphi)$—specifically, its connection to the formula φ—by requiring that $do(\varphi)$ correspond to a map from Ω to Ω such that $\omega \mapsto c(\omega, \varphi)$ for a suitable selection function c. One of the novel aspects of our proof consists in showing how to determine the selection function from preferences on acts.

Since our framework can be viewed a specialization of the BEH framework (with our actions having additional, language-based structure as described), rather than proving our representation theorem from scratch, we can reuse much of their construction. Thus, we will present the same axioms (adapted to our notation) that BEH present, and subseqently augment them with new principles that allow us to construct the selection function.

3.1 Cancellation

BEH's main axiom is a *cancellation law*. Explaining this requires a few preliminary definitions, beginning with the notion of a *multiset*, which can be thought of as a set that allows for multiple instances of each of its elements; two multisets are equal just in case they contain the same elements *with the same multiplicities*. For example, the multiset $\{\!\{a,a,a,b,b\}\!\}$ is different from the multiset $\{\!\{a,b,b,b,b\}\!\}$: both multisets have five elements, but the mulitiplicity of a and b differ.

Given any subset $X \subseteq \Phi$, let $\varphi_X = \bigwedge_{p \in X} p \wedge \bigwedge_{q \notin X} \neg q$. Intuitively, φ_X is a "complete description" of the truth values of all primitive propositions in the language $\mathcal{L}(\Phi)$, namely the description that says for each primitive proposition p that it is true iff it belongs to X. An **atom** is any formula of the form φ_X. Since $\mathcal{L}(\Phi)$ is a propositional language and we use classical semantics for propositional logic, for all formulas $\varphi \in \mathcal{L}(\Phi)$ and atoms φ_X, the truth of φ is determined by φ_X: either $\models \varphi_X \to \varphi$, or $\models \varphi_X \to \neg\varphi$.

It is therefore not surprising that every action in $\alpha \in \mathcal{A}_F$ can be identified with a function $f_\alpha : 2^\Phi \to F$, defined recursively as follows:

$$\begin{aligned} f_{do(\varphi)}(X) &= \varphi \\ f_{\textbf{if } \psi \textbf{ then } \alpha \textbf{ else } \beta}(X) &= \begin{cases} f_\alpha(X) & \text{if } \models \varphi_X \to \psi \\ f_\beta(X) & \text{if } \models \varphi_X \to \neg\psi. \end{cases} \end{aligned}$$

BEH define atoms in the same way and use them to define functions from atoms to primitive choices just as we did above (replace $do(\varphi)$ by an arbitrary primitive choice).[4]

Now we can state the central cancellation law that enables us to apply the BEH representation theorem:

(Canc) Let $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \in \mathcal{A}_F$, and suppose that for each $X \subseteq \Phi$ we have $\{\!\{f_{\alpha_1}(X), \ldots, f_{\alpha_n}(X)\}\!\} = \{\!\{f_{\beta_1}(X), \ldots, f_{\beta_n}(X)\}\!\}$. Then, if for all $i < n$ we have $\alpha_i \succeq \beta_i$, it follows that $\beta_n \succeq \alpha_n$.

Intuitively, this says that if we get the same collection of outcomes with $\alpha_1, \ldots, \alpha_n$ as with $\beta_1, \ldots, \beta_n$ (taking multiplicity into account) in each state, then we should view the collection $\{\!\{\alpha_1, \ldots, \alpha_n\}\!\}$ and $\{\!\{\beta_1, \ldots, \beta_n\}\!\}$ as equally good. Thus, if α_i is at least as good as β_i for $i = 1, \ldots, n-1$, then, to balance things out, β_n should be at least as good as α_n.

As pointed out by BEH, Cancellation is a surprisingly powerful axiom. In particular, BEH show that we can use **(Canc)** to derive many simpler (and more classical) principles of choice: that $\succeq$ is reflexive and transitive, that *independence* holds,[5] and that if α and β are *equivalent* in the sense that $f_\alpha = f_\beta$, then $\alpha \sim \beta$. (However, it should be noted that Cancellation seems stronger than the conjunection of these axioms.)

3.2 Selection axioms

To present the new axioms that will allow us to construct an appropriate selection function as part of the representation theorem, it will be helpful to introduce some new notation. To begin, we write **if** φ **then** α as a shorthand for **if** φ **then** α **else** $do(true)$. Intuitively, the action $do(true)$ corresponds to doing "nothing", since *true* is true no matter what, so we might think of "otherwise nothing" as being the default in case no explicit **else...** clause is given. Of course, for this to make sense we must have $true \in F$; we make this assumption henceforth.

Next we define an abbreviation for *conditional preference*, familiar from Savage's classical development [12]: write $\alpha \succeq_\varphi \beta$ as an abbreviation for (**if** φ **then** α) $\succeq$ (**if** φ **then** β).[6] When $\varphi = \varphi_X$, we write $\alpha \succeq_X \beta$ for $\alpha \succeq_{\varphi_X} \beta$, and we extend this notation to strict conditional preference and conditional indifference in the obvious way.

Our first axiom is related to the centering constraint for selection functions (i.e., that if φ is true at a state, then that state automatically counts the "closest" φ-state):

[4] Techncially, we are not mapping atoms to primitive acts, but since there is an obvious bijection $X \mapsto \varphi_X$ beween sets of primitive proposition and atoms, and an obvious bijection $\varphi \mapsto do(\varphi)$ between elements of F and primitive acts, we really can be thought of as doing just that.

[5] That is, for all $\alpha, \beta, \gamma, \gamma' \in \mathcal{A}_F$ and all $\varphi \in F$,

$$(\textbf{if } \varphi \textbf{ then } \alpha \textbf{ else } \gamma \succeq \textbf{if } \varphi \textbf{ then } \beta \textbf{ else } \gamma) \Leftrightarrow (\textbf{if } \varphi \textbf{ then } \alpha \textbf{ else } \gamma' \succeq \textbf{if } \varphi \textbf{ then } \beta \textbf{ else } \gamma').$$

[6] As BEH show, the cancellation law implies independence, so in fact we have $\alpha \succeq_\varphi \beta$ iff for *all* γ, **if** φ **then** α **else** $\gamma \succeq$ **if** φ **then** β **else** γ.

(Cent) If $\models \psi \to \varphi$, then (**if** ψ **then** $do(\varphi)) \sim do(true)$.

To build intuition it's helpful to consider the special case where $\psi = \varphi$, in which case **(Cent)** just says that doing φ precisely when φ is already the case (and otherwise doing nothing) is the same as doing nothing. Here of course by "the same" what is really meant is that the agent is indifferent between those two acts. Since we are trying to bootstrap properties of a selection function from the agent's preferences, all our principles will ultimately need to bottom out in statements about what the agent does or does not have a preference between. The general statement of **(Cent)** simply expands this reasoning to cases where the condition ψ entails the result of the action, φ, and so again in this case $do(\varphi)$ happens only in cases where φ is already true.

Lemma 1. *If (M,c) is a selection model, c satisfies centering, and $\models \psi \to \varphi$, then*

$$[\![\textbf{\textit{if }}\psi \textbf{\textit{ then }} do(\varphi)]\!]_{M,c} = id_\Omega = [\![do(true)]\!]_{M,c}.$$

Our second axiom is meant to capture the idea that *sufficiently specific conditions* resolve any ambiguity (expressible in the underlying language) about the effect of an action:

(SSC) If $\models \varphi \leftrightarrow (\varphi_1 \vee \cdots \vee \varphi_n)$, then $\forall X \subseteq \Phi$, $\exists i \in \{1,\dots,n\}$ such that for all ψ satisfying $\models \varphi_i \to \psi$ and $\models \psi \to \varphi$, we have $do(\psi) \sim_X do(\varphi_i)$.

This requires some unpacking. As above, it is illuminating to begin by considering the special case where $\psi = \varphi$. Then $\models \psi \to \varphi$ holds trivially and $\models \varphi_i \to \psi$ is true by assumption, so we can read **(SSC)** intuitively as follows: If φ is ambiguous between a variety of (potentially) more precise statements (namely, $\varphi_1,\dots,\varphi_n$), then for any sufficiently specific condition (i.e., any atom φ_X), there is at least one precisification φ_i of φ such that, conditional on φ_X, doing φ is equivalent to doing φ_i (from the agent's perspective).

This, as well as the more general statement of **(SSC)**, follows from the assumption that the selection function c is induced by a language-based family of well-orders.

Lemma 2. *If (M,c) is a selection model where c is induced by the well-orders $\leq = \{\leq_\omega : \omega \in \Omega\}$, $\leq$ is language-based, $\models \varphi \leftrightarrow (\varphi_1 \vee \cdots \vee \varphi_n)$, and $X \subseteq \Phi$, then $\exists i \in \{1,\dots,n\}$ such that for all ψ satisfying $\models \varphi_i \to \psi$ and $\models \psi \to \varphi$ and all $\omega \in [\![\varphi_X]\!]$, we have $[\![do(\psi)]\!]_{M,c}(\omega) = [\![do(\varphi_i)]\!]_{M,c}(\omega)$.*

The next idea is crucial to the ultimate construction of our selection function. For each atom φ_W, we will define a total preorder[7] $\sqsubseteq_W$ on the set of atoms that will in turn be extended to a linear order and used to specify the selection function. Formally, we define:

$$\varphi_X \sqsubseteq_W \varphi_Y \text{ iff } do(\varphi_X \vee \varphi_Y) \sim_W do(\varphi_X).$$

Loosely speaking, $\varphi_X \sqsubseteq_W \varphi_Y$ says that in φ_W-states, the ambiguity inherent in doing $\varphi_X \vee \varphi_Y$ is resolved in the agent's mind in favour of doing φ_X; this is why the agent is indifferent (conditional on φ_W) between doing $\varphi_X \vee \varphi_Y$ and just doing φ_X. In this sense we think of φ_X as being at least as "close" to φ_W as φ_Y is.

Note that the definition above requires F to contain all atoms as well as all pairwise disjunctions of atoms. This richness in F is what allows us to use the agent's preferences on actions to define an appropriate preorder. We make this assumption henceforth. It is an interesting question to what extent the ensuing construction can be carried out without this assumption; we return to this point in Section 4.

Now we can state our third axiom, which simply says that this notion of closeness is transitive:

(Trans) For all $W,X,Y,Z \subseteq \Phi$, if $\varphi_X \sqsubseteq_W \varphi_Y$ and $\varphi_Y \sqsubseteq_W \varphi_Z$, then $\varphi_X \sqsubseteq_W \varphi_Z$.

[7] A *total preorder* is a complete and transitive relation (so, unlike a linear order, it need not be antisymmetric).

Lemma 3. **(SSC)** *implies that each* $\sqsubseteq_W$ *is complete.*

Lemma 4. *If* **(SSC)** *and* **(Trans)** *hold, then each* $\sqsubseteq_W$ *is a total preorder and can be extended to a well-order* $\leq_W$ *on the set of atoms; if, in addition,* **(Cent)** *holds, then each* $\leq_W$ *can be defined so that* φ_W *is the* $\leq_W$*-minimal element.*

Given a family of well-orders $\{\leq_W : W \subseteq \Phi\}$ as defined in Lemma 4, let $min_{\leq}(W,\varphi)$ denote the unique $X \subseteq \Phi$ such that φ_X is $\leq_W$-minimal in $\{\varphi_Y : \models \varphi_Y \rightarrow \varphi\}$. So φ_X is the "closest" atom compatible with φ to φ_W; intuitively, then, doing φ in a φ_W situation should essentially amount to doing φ_X. This is precisely what the next lemma asserts.

Lemma 5. *If* **(SSC)** *and* **(Trans)** *hold, then* $do(\varphi) \sim_W do(\varphi_{min_{\leq}(W,\varphi)})$.

3.3 The representation theorem

Theorem 1. *If* $\succeq$ *is a complete binary relation on* $\mathcal{A}_F$ *satisfying* **(Canc)**, **(Cent)**, **(SSC)**, *and* **(Trans)**, *then there is a language-based SEU representation* $\succeq$.

Proof. We begin by following the proof in [3, Theorem 2] to obtain a state-dependent representation with state space 2^Φ and outcome space F.[8] More precisely, we consider the set of functions $\mathcal{F} = \{f_\alpha : \alpha \in \mathcal{A}_F\}$ defined in Section 3.1, which can be viewed as Savage acts in the classical sense [12]. The relation $\succeq$ on $\mathcal{A}_F$ induces a relation $\succeq^*$ on $\mathcal{F}$ defined as follows:

$$f_\alpha \succeq^* f_\beta \Leftrightarrow \alpha \succeq \beta.$$

As discussed, **(Canc)** implies that $\alpha \sim \alpha'$ whenever $f_\alpha = f_{\alpha'}$, so $\succeq^*$ is well-defined; moreover, as BEH show, **(Canc)** is strong enough to yield the desired state-dependent representation result for $\succeq^*$, namely, that there exists a function $u^* : 2^\Phi \times F \rightarrow \mathbb{R}$ such that, for all $f, g \in \mathcal{F}$,

$$f \succeq^* g \Leftrightarrow \sum_{X \in 2^\Phi} u^*(X, f(X)) \geq \sum_{X \in 2^\Phi} u^*(X, g(X)).$$

Up to now we have mirrored the proof given by BEH exactly, which has given us a utility function u^* but also an outcome space that we don't want. Moreover, the utility function is *state-dependent*; it takes as arguments both a state and an outcome. We want a utility function that depends only on states (which for us are the same as outcomes). Thus, our task now is to transform this result into a selection model that we can use to give a language-based SEU representation of $\succeq$ (including a utility function defined only on states).

Set $\Omega = 2^\Phi \times 2^\Phi$; so our state space is isomorphic to *pairs* of atoms. This is a technical maneuver that allows us to "factor out" probabilities from the state-dependent utility function u^* we already have. Loosely speaking, given $(X,Y) \in \Omega$, the first component X represents how things are, while the second component Y represents how things *were*. This intuition should become clearer as we continue.

We define a basic model $M = (\Omega, [\![\cdot]\!]_M)$ by specifying the valuation on Ω as follows:

$$[\![p]\!]_M = \{(X,Y) \in \Omega : \models \varphi_X \rightarrow p\}.$$

In other words, p is true at (X,Y) just in case φ_X entails p. Note that the valuation only depends on the first component X of the state (X,Y).

[8] "State-dependent" here means that the utility function constructed will depend not only on outcomes but on states as well.

Next we specify a parametrized family of well-orders on Ω that we can use to induce a selection function. First define

$$(X,X') \sqsubseteq_{W,W'} (Y,Y') \text{ iff } \varphi_X \leq_W \varphi_Y.$$

Again, we are ignoring the second component. This is clearly a well-order when restricted to the first component of the state space, but not in general, since by definition we have $(X,X') \sqsubseteq_{W,W'} (Y,Y')$ and $(Y,Y') \sqsubseteq_{W,W'} (X,X')$ whenever $X = Y$. However, as usual, we can extend these relations to well-orders $\leq_{W,W'}$ on all of Ω simply by choosing a linear order for each set of the form $\Omega_X := \{(X,Y) : Y \in 2^\Phi\}$, and in so doing we can insist that for each fixed X, the state (X,W) is $\leq_{W,W'}$-minimal on the set Ω_X.

This is the first time we have paid attention to the second component of the state. Roughly speaking, we are ensuring that the order $\leq_{W,W'}$ "remembers" the set W. More perspicuously, it is easy to see that if c is the selection function induced by the family $\{\leq_{(W,W')} : (W,W') \in \Omega\}$, then for each $(W,W') \in \Omega$ and all $\varphi \in \mathcal{L}$, we have

$$[\![do(\varphi)]\!]_{M,c}(W,W') = c((W,W'),\varphi) = (min_{\leq}(W,\varphi),W). \tag{2}$$

That is, the closest φ-state to (W,W') encodes both the closest atom compatible with φ to φ_W (in the first component) *and* the state W that we started from (in the second component).

Now we can define our utility function and probability measure. Let π be any probability measure on Ω satisfying $\pi(\Omega_X) > 0$ for all X. Next, define $u : \Omega \to \mathbb{R}$ by

$$u(X,W) = \frac{u^*(W,\varphi)}{\pi(\Omega_W)}, \text{ for some } \varphi \text{ such that } min_{\leq}(W,\varphi) = X.$$

Of course, we need to check that u is well-defined, and we do so in Lemma 6. But first some intuition is in order. Thinking back to the state-dependent utility function u^*, a reasonable first gloss of the meaning of $u^*(W,\varphi)$ might be "the utility of doing φ in W".[9] The point is that u^* is specifying the utility value not of an action in itself or the "result" of an action, but rather the result of an action *if you started in a certain state*. This is all very informal, but the idea is just to provide some intuition for why, in defining our utility function u from u^*, we need to appeal to a rich enough notion of state that can "remember" what the "previous" state was—intuitively, the state we were at before the action was performed.

Lemma 6. *The function u is well-defined.*

The last thing we need to show is that the selection model (M,c) we have built, along with π and u, gives us an expected utility representation of $\succeq$. So let $\alpha,\beta \in \mathcal{A}_F$ and suppose that $\alpha \succeq \beta$. By definition this is equivalent to $f_\alpha \succeq^* f_\beta$, which by the state-dependent representation result is in turn equivalent to

$$\sum_{W \in 2^\Phi} u^*(W, f_\alpha(W)) \geq \sum_{W \in 2^\Phi} u^*(W, f_\beta(W)). \tag{3}$$

Now observe that, for each $W \in 2^\Phi$,

$$\begin{aligned}
u^*(W, f_\alpha(W)) &= \pi(\Omega_W) \cdot u(min_{\leq}(W, f_\alpha(W)), W) && \text{(by definition of } u\text{)} \\
&= \pi(\Omega_W) \cdot u([\![do(f_\alpha(W))]\!]_{M,c}(W,W')) && \text{(from (2))} \\
&= \pi(\Omega_W) \cdot u([\![\alpha]\!]_{M,c}(W,W')) && \text{(by definition of } f_\alpha \text{ and } (M,c)\text{)}.
\end{aligned}$$

[9] Though this isn't quite right—it's more like the product of that utility with the probability of W, which is why we have to factor that probability out in defining our utility function.

Note that in the above W' can be *any* element of 2^Φ, since it's not taken into account in determining the result of an action. That means we can rewrite the above as

$$u^*(W, f_\alpha(W)) = \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u([\![\alpha]\!]_{M,c}(W, W')).$$

Of course, an analogous equation holds for $u^*(W, f_\beta(W))$. Thus, (3) is equivalent to:

$$\sum_{W \in 2^\Phi} \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u([\![\alpha]\!]_{M,c}(W, W')) \geq \sum_{W \in 2^\Phi} \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u([\![\beta]\!]_{M,c}(W, W')),$$

which is exactly the right-hand side of (1), completing the proof. □

4 Discussion

We have considered a framework in which both the conditions for and the results of an action are given by simple descriptions in a fixed language. These descriptions may not be maximally specific, so the results of actions can be underspecified and therefore "open to interpretation". We have shown that, in this context, agents whose preferences satisfy certain constraints can be represented as if they are expected utility maximizers who interpret each underspecified action using a selection function identical to that employed in standard semantics for counterfactual conditionals.

The representation theorem presented in this extended abstract might be viewed as a sort of "proof of concept", namely, that such representation results are possible and even natural. This opens the door for a variety of related results connecting different assumptions about the selection function to different constraints on the agent's preferences. As we mentioned above, there are a number of standard assumptions along these lines in the literature on counterfactuals.

The underlying language we chose to work with can also be altered. Perhaps most obviously, we might consider allowing countably-many primitive propositions. In this case, we cannot straightforwardly use atoms as the basis for the state space in the representation theorem, and in general we might need to relax the notion of a "complete description" to something like a "sufficiently detailed description". Going in the other direction, we might also considering dropping some of the richness constraints we imposed. For instance, we assumed that F contains all atoms (and all pairwise disjunctions of atoms). Can this assumption be relaxed?

In our framework, because we use the same descriptions for both states and outcomes, we found it convenient to identify the two. This in turn makes it straightforward to extend to a richer language of acts, where we allow *sequential actions*, implemented directly by function composition. That is, we can allow actions of the form $do(\varphi);do(\psi)$ ("first do φ, then do ψ"), or more generally, $\alpha;\beta$. Thus, the (underspecified!) results of the first action are directly relevant to the conditions under which the second action is executed, which may allow for entirely new and intriguing ways of encoding modeling features via constraints on preferences.

Finally, generalizing this framework to multiple agents is of interest. Indeed, the original motivation for this work is doubly relevant in multi-agent settings: two different decision-makers might conceive of the same action in different ways, by associating it with different functions. For example, we should be able to model two agents who agree about their values and have the same beliefs about the likelihoods of uncertain events, but still have different preferences over actions—intuitively, because they interpret the "default" way of implementing actions differently (in other words, they have the same utility function and probability measure, but different selection functions).

In short, this area is ripe for further exploration, with many theoretical and practical applications.

A Proofs

Lemma 1. *If (M,c) is a selection model, c satisfies centering, and $\models \psi \to \varphi$, then*

$$[\![\textbf{\textit{if}}\ \psi\ \textbf{\textit{then}}\ do(\varphi)]\!]_{M,c} = id_\Omega = [\![do(true)]\!]_{M,c}.$$

Proof. By definition, we have

$$\begin{aligned}[\![\textbf{if}\ \psi\ \textbf{then}\ do(\varphi)]\!]_{M,c}(\omega) &= \begin{cases} [\![do(\varphi)]\!]_{M,c}(\omega) & \text{if } \omega \in [\![\psi]\!] \\ [\![do(true)]\!]_{M,c}(\omega) & \text{if } \omega \notin [\![\psi]\!]. \end{cases} \\ &= \begin{cases} c(\omega,\varphi) & \text{if } \omega \in [\![\psi]\!] \\ c(\omega,true) & \text{if } \omega \notin [\![\psi]\!]. \end{cases}\end{aligned}$$

But since $[\![\psi]\!] \subseteq [\![\varphi]\!]$ by assumption, in either case, centering applies and guarantees that

$$[\![\textbf{if}\ \psi\ \textbf{then}\ do(\varphi)]\!]_{M,c}(\omega) = \omega. \qquad \square$$

Lemma 2. *If (M,c) is a selection model where c is induced by the well-orders $\leq = \{\leq_\omega : \omega \in \Omega\}$, $\leq$ is language-based, $\models \varphi \leftrightarrow (\varphi_1 \vee \cdots \vee \varphi_n)$, and $X \subseteq \Phi$, then $\exists i \in \{1,\ldots,n\}$ such that for all ψ satisfying $\models \varphi_i \to \psi$ and $\models \psi \to \varphi$ and all $\omega \in [\![\varphi_X]\!]$, we have $[\![do(\psi)]\!]_{M,c}(\omega) = [\![do(\varphi_i)]\!]_{M,c}(\omega)$.*

Proof. Let $\omega \in [\![\varphi_X]\!]$ and choose i such that $c(\omega,\varphi) \in [\![\varphi_i]\!]$. This is possible since we know $c(\omega,\varphi) \in [\![\varphi]\!]$ and, by assumption, $[\![\varphi]\!] = [\![\varphi_1]\!] \cup \ldots \cup [\![\varphi_n]\!]$. Since $c(\omega,\varphi)$ is the $\leq_\omega$-minimal element of $[\![\varphi]\!]$, it follows that for any set T with $c(\omega,\varphi) \in T \subseteq [\![\varphi]\!]$, $c(\omega,\varphi)$ is also the $\leq_\omega$-minimal element of T. In particular, since $c(\omega,\varphi) \in [\![\varphi_i]\!] \subseteq [\![\psi]\!] \subseteq [\![\varphi]\!]$, this implies that $c(\omega,\varphi)$ is the $\leq_\omega$-minimal element of both $[\![\varphi_i]\!]$ and $[\![\psi]\!]$. Thus, by definition, $c(\omega,\varphi_i) = c(\omega,\psi)$, so

$$[\![do(\psi)]\!]_{M,c}(\omega) = c(\omega,\psi) = c(\omega,\varphi_i) = [\![do(\varphi_i)]\!]_{M,c}(\omega).$$

Since $\omega \models \varphi_X$ and this completely determines the theory of ω, we know that for any other $\omega' \in [\![\varphi_X]\!]$, $\omega' \equiv \omega$, so $c(\omega',\varphi) \equiv c(\omega,\varphi)$. This guarantees that $c(\omega',\varphi) \in [\![\varphi_i]\!]$; in other words, the same choice of i works for all states in $[\![\varphi_X]\!]$, which completes the proof. $\square$

Lemma 3. **(SSC)** *implies that each $\sqsubseteq_W$ is complete.*

Proof. Fix any two atoms φ_X and φ_Y. We apply **(SSC)** in the case where $\varphi = \varphi_X \vee \varphi_Y$, $\varphi_1 = \varphi_X$, $\varphi_2 = \varphi_Y$, and $\psi = \varphi$. Then we know that given any $W \subseteq \Phi$, either $do(\varphi) \sim_W do(\varphi_1)$ or $do(\varphi) \sim_W do(\varphi_2)$, that is, either $do(\varphi_X \vee \varphi_Y) \sim_W do(\varphi_X)$ or $do(\varphi_X \vee \varphi_Y) \sim_W do(\varphi_Y)$, which established completeness. $\square$

Lemma 4. *If* **(SSC)** *and* **(Trans)** *hold, then each $\sqsubseteq_W$ is a total preorder and can be extended to a well-order $\leq_W$ on the set of atoms; if, in addition,* **(Cent)** *holds, then each $\leq_W$ can be defined so that φ_W is the $\leq_W$-minimal element.*

Proof. The fact that $\sqsubseteq_W$ is a total preorder follows immediately from **(Trans)** and Lemma 3. Moreover, it is easy to see that any total preorder on a finite set can be extended to a well-order (by choosing an arbitrary linear order for each subset of $\sqsubseteq_W$-equivalent atoms). To see that this can be done in such a way that φ_W is the $\leq_W$-minimal element, it suffices to show that for every $X \subseteq \Phi$, we have $\varphi_W \sqsubseteq_W \varphi_X$, or in other words, $do(\varphi_W \vee \varphi_X) \sim_W do(\varphi_W)$. The result now follows from two applications of **(Cent)**. First we apply it in the case where $\psi = \varphi_W$ and $\varphi = \varphi_W \vee \varphi_X$ to obtain $(\textbf{if}\ \varphi_W\ \textbf{then}\ do(\varphi_W \vee \varphi_X)) \sim do(true)$;

then we apply it in the case where $\psi = \varphi = \varphi_W$ to obtain (**if** φ_W **then** $do(\varphi_W)) \sim do(true)$. Transitivity of $\sim$ therefore yields

$$(\textbf{if } \varphi_W \textbf{ then } do(\varphi_W \vee \varphi_X)) \sim (\textbf{if } \varphi_W \textbf{ then } do(\varphi_W)),$$

which by definition is equivalent to $do(\varphi_W \vee \varphi_X) \sim_W do(\varphi_W)$. □

Lemma 5. *If* **(SSC)** *and* **(Trans)** *hold, then* $do(\varphi) \sim_W do(\varphi_{min_{\leq}(W,\varphi)})$.

Proof. Let $X = min_{\leq}(W,\varphi)$, and let $\varphi_{X_1}, \ldots, \varphi_{X_n}$ enumerate all the atoms compatible with φ. Then by definition we know that $X = X_j$ for some j. We also clearly have $\models \varphi \leftrightarrow (\varphi_{X_1} \vee \cdots \vee \varphi_{X_n})$, so we can apply **(SSC)** (taking $\psi = \varphi$) to find an i such that $do(\varphi) \sim_W do(\varphi_{X_i})$.

By definition of X, we know that $\varphi_X \leq_W \varphi_{X_i}$, which means $do(\varphi_X \vee \varphi_{X_i}) \sim_W do(\varphi_X)$. On the other hand, since $\models \varphi_{X_i} \rightarrow (\varphi_X \vee \varphi_{X_i})$ and $\models (\varphi_X \vee \varphi_{X_i}) \rightarrow \varphi$, **(SSC)** also tells us (taking $\psi = \varphi_X \vee \varphi_{X_i}$ this time) that $do(\varphi_X \vee \varphi_{X_i}) \sim_W do(\varphi_{X_i})$. By transitivity of $\sim_W$ we therefore have $do(\varphi_{X_i}) \sim_W do(\varphi_X)$, and therefore $do(\varphi) \sim_W do(\varphi_X)$, as desired. □

Lemma 6. *The function u is well-defined.*

Proof. What we need to show that is that if $min_{\leq}(W,\varphi) = X$ and also $min_{\leq}(W,\varphi') = X$, then $u^*(W,\varphi) = u^*(W,\varphi')$. By Lemma 5, we know that $do(\varphi) \sim_W do(\varphi_X)$, and also that $do(\varphi') \sim_W do(\varphi_X)$. Focusing on the first of these two indifferences to begin with, by definition we have

$$\textbf{if } \varphi_W \textbf{ then } do(\varphi) \sim \textbf{if } \varphi_W \textbf{ then } do(\varphi_X).$$

Setting $\alpha =$ **if** φ_W **then** $do(\varphi)$ and $\beta =$ **if** φ_W **then** $do(\varphi_X)$, it follows that $f_\alpha \sim^* f_\beta$ (by definition of $\succeq^*$). Thus, from the state-dependent representation result, we can deduce that

$$\sum_{Z \in 2^\Phi} u^*(Z, f_\alpha(Z)) = \sum_{Z \in 2^\Phi} u^*(Z, f_\beta(Z)).$$

But it's easy to see that whenever $Z \neq W$, $f_\alpha(Z) = f_\beta(Z)$, so we can cancel all those terms in the equality above to arrive at $u^*(W, f_\alpha(W)) = u^*(W, f_\beta(W))$. This yields $u^*(W,\varphi) = u^*(W,\varphi_X)$, since clearly $f_\alpha(W) = \varphi$ and $f_\beta(W) = \varphi_X$. Analogous reasoning starting from the fact that $do(\varphi') \sim_W do(\varphi_X)$ leads us to $u^*(W,\varphi') = u^*(W,\varphi_X)$. Putting these together gives $u^*(W,\varphi) = u^*(W,\varphi')$, as desired. □

References

[1] D. Ahn (2007): *Ambiguity without a state space*. Forthcoming, *Review of Economic Studies*.

[2] D. Ahn & H. Ergin (2007): *Framing contingencies*. Unpublished manuscript.

[3] L. E. Blume, D. Easley & J. Y. Halpern (2006): *Redoing the Foundations of Decision Theory*. In: *Principles of Knowledge Representation and Reasoning: Proc. Tenth International Conference (KR '06)*, pp. 14–24. A longer version, entitled "Constructive decision theory", can be found at http://www.cs.cornell.edu/home/halpern/papers/behfinal.pdf.

[4] E. Dekel, B. Lipman & A. Rustichini (2001): *Representing preferences with a unique subjective state space*. *Econometrica* 69, pp. 891–934.

[5] P. Ghirardato (2001): *Coping with ignorance: unforeseen contingencies and non-additive uncertainty*. *Economic Theory* 17, pp. 247–276.

[6] I. Gilboa & D. Schmeidler (2004): *Subjective distributions*. *Theory and Decision* 56, pp. 345–357.

[7] E. Karni (2006): *Subjective expected utility theory without states of the world*. Journal of Mathematical Economics 42, pp. 325–342.

[8] D. Kreps (1992): *Static choice and unforeseen contingencies*. In P. Dasgupta, D. Gale & O. Hart, editors: Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn, MIT Press, Cambridge, MA.

[9] B. L. Lipman (1999): *Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality*. Review of Economic Studies 66, pp. 339–361.

[10] M. Machina (2006): *States of the World and State of Decision Theory*. In D. Meyer, editor: The Economics of Risk, W. E. Upjohn Institute.

[11] J. Pearl (1995): *Causal diagrams for empirical research*. Biometrika 82(4), pp. 669–710.

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

[13] R. C. Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: Studies in Logical Theory, Blackwell, Oxford, U.K., pp. 98–112.

[14] A. Tversky & D. J. Koehler (1994): *Support theory: A nonextensional representation of subjective probability*. Psychological Review 101(4), pp. 547–567.