J Philos Logic (2017) 46:467–506
DOI 10.1007/s10992-016-9408-8

A Representation Theorem for Frequently Irrational Agents

Edward Elliott[1]

Received: 26 December 2015 / Accepted: 4 August 2016 / Published online: 19 August 2016

Abstract The standard representation theorem for expected utility theory tells us that if a subject's preferences conform to certain axioms, then she can be represented as maximising her expected utility given a particular set of credences and utilities—and, moreover, that having those credences and utilities is the *only* way that she could be maximising her expected utility (given her preferences). However, the kinds of agents these theorems seem apt to tell us anything about are highly idealised, being (amongst other things) always probabilistically coherent with infinitely precise degrees of belief and full knowledge of all *a priori* truths. Ordinary subjects do not look very rational when compared to the kinds of agents usually talked about in decision theory. In this paper, I will develop an expected utility representation theorem aimed at the representation of those who are neither probabilistically coherent, logically omniscient, nor expected utility maximisers across the board—that is, agents who are *frequently irrational*. The agents in question may be deductively fallible, have incoherent credences, limited representational capacities, and fail to maximise expected utility for all but a limited class of gambles.

Keywords Representation theorem · Preferences · Credences · Utilities · Irrational

✉ Edward Elliott
edd.elliott@gmail.com

[1] School of Philosophy, Religion and History of Science, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK

1 Introduction

The standard expected utility representation theorem tells us roughly that if a subject's preferences conform to certain axioms, then she can be represented as maximising her expected utility under a particular set of credences and utilities—and, moreover, that given the facts about her preferences, having those credences and utilities is the *only* way that she could be an expected utility maximiser.

It has long been thought that representation theorems can tell us something interesting about the connection between a subject's preferences on the one hand, and her credences and utilities on the other. Some go for the relatively weak thesis that representation theorems establish a clear evidential connection between the former states and the latter (e.g., [45, pp. 60-61]). This is not especially controversial, though we do need to be careful on the details of just how strong that evidential connection can be taken to be. From the beginning, however, many have also thought to use representation theorems in support of the more philosophically interesting thesis that preferences are metaphysically more fundamental than credences and utilities, and that the latter can be put towards characterising the former.[1] In this paper, I will be mainly interested in the metaphysical thesis, though what I say also has straightforward connections to the weaker evidential thesis.

In Section 2, I will explore how the relevant kind of metaphysical connections might be made and motivated, and highlight some key conditions that any representation theorem should satisfy if it is to be fruitfully put to such purposes. I will also note that contemporary theorems fall short of these conditions in a number of respects. At best, the kinds of agents that these theorems seem apt to tell us much about are highly idealised: they never make the wrong decisions, they have perfectly probabilistically coherent credences, full knowledge of all epistemically necessary truths and logical equivalences, and infinitely precise opinions for any proposition that they consider.[2]

The average person on the street does not look very rational when compared to the hyperrational angels for whom these theorems are appropriate (i.e., the kind of agent who both satisfies the preference axioms of expected utility theory *and* plausibly has credences and utilities accurately represented by the corresponding rational credence and utility functions). We *ordinary agents* manage to get by, of course, and we seem to generally act in such a way as to tend to bring about the kinds of things we desire given the way we take the world to be. But we're not even *close* to ideally rational

[1] Ramsey [34] developed the first expected utility representation theorem, which he intended as the basis for a definition of credences and utilities. Authors sympathetic to the metaphysical application of representation theorems include Cozic and Hill [8], Davidson [9, 10], Ells [13], Harsanyi [18], Jeffrey [20], Maher [29, 30], and Pettit [32, pp. 171–172]. Note that the issue here is not whether credences and utilities just *are* preference states, nor whether they are reducible to preferences *alone*; these are much stronger claims than we need commit ourselves to. See Section 2 for discussion.

[2] Here and throughout, I will use 'epistemically necessary'and 'epistemically possible' (or sometimes just 'necessary', 'possible') in more or less the sense explicated by Chalmers [6, 7]. Essentially: P is epistemically possible iff it can't be ruled out *a priori*, and epistemically necessary iff it is *a priori* knowable.

in the way that, for instance, full probabilistic coherence would require. It would be nice if we could have a representation theorem for *us*, too.[3]

In Sections 3 and 4, I develop a theorem aimed at the representation of even very frequently irrational agents. By 'frequently irrational agents', I mean those who:

a) May be probabilistically incoherent by, for example, having non-additive and non-monotonic credences, and less-than-full belief in some *a priori* truths;
b) May have credences and utilities towards only a set of relatively non-specific propositions (which need not have an algebraic structure);
c) May not consistently maximise expected utility.

Those mainly interested in the technical results might skim over Section 2 without loss of comprehension. The theorem to be established in Section 3 is a modified version of the one developed in [15], itself based on the work of Ramsey [34]. The central difference is a generalised approach to the relation that must hold between propositions used in characterising the space of options (with some corresponding new and modified axioms to accommodate). This allows for interesting re-interpretations of the theorem's axioms, and, as a consequence of this, much more permissive representations of credences and utilities. The developments discussed in Section 4 are new, and involve the removal or modification of several problematic aspects of Section 3's theorem.

Section 5 concludes with a discussion on the plausibility of the axioms underlying the new theorem, and a comparison between the basic formal structures it uses and those employed by other prominent representation theorems. In some important respects, the formal basis for the theorem developed herein is significantly different than the better known theorems of Savage [36] and Jeffrey [21]. Most importantly, the axioms put forward here don't say anything about what our subject's preferences have to be like *in general*; they are only intended to apply to a fragment of her overall preference structure. This contributes to the flexibility of its representation of credences and utilities, though it also limits the scope with which the theorem can be used explain and make predictions about our preferences overall.

2 Functionalism and the Graded Attitudes

I take it for granted that ordinary agents have *preferences* regarding different ways the world might be, and that these preferences are a kind of comparative propositional attitude. I also assume that ordinary agents have *credences* and *utilities*, which are also best construed as propositional attitudes. Finally, I assume that it makes sense to numerically measure degrees of confidence and desire, though the exact shape that those measures should take is something that should be left up for grabs.

[3]Representation theorems for non-expected utility often forego probability functions in favour of non-additive Choquet capacities, Dempster-Shafer belief and plausibility functions, sets of probability functions, and so on. These models tend to be somewhat more realistic, but only marginally so—e.g., each implies that if P necessitates Q, then $Cr(Q) \geq Cr(P)$.

In speaking of credences and utilities, I mean to refer to the ordinary folk psychological notions of graded belief and desire. There are also purely stipulative senses of 'credence' and 'utility', whereby they are supposed to be high-level theoretical constructs designed to relate and explain choice behaviour, with no deep connections to folk psychology or everyday attitude attributions. These are not my topic. Gradation is an important and, I expect, ineliminable part of the folk conception of the mind. Inasmuch as we are willing to accept that ordinary agents have propositional attitudes at all, we can take it for granted that at least some of these attitudes come in degrees. As theorists, there is some room for explication and systematisation, but we ought not deviate too far from the ordinary concepts lest we change the topic.

Relatedly, I do not assume that facts about preferences are reducible to facts about choices. There are plenty of possibilities that we might like to have a choice between than we will ever actually have a chance to—preferences which may never factor into any decision we have to make. So, our preferences are not mere re-descriptions of our actual choice behaviour. Nor do I think it immediately obvious that facts about preferences might be reducible to facts about our *dispositions* to choose under various kinds of counterfactual circumstance (for discussion, see [29], pp. 14–15). Choice dispositions provide evidence for preferences, but the preferences themselves are better thought of as causally antecedent mental states that are in principle separable from our choices and choice dispositions.

It should not be taken for granted that agents have preferences regarding *every* way the world may be. Many possibilities may be too fine-grained to even contemplate, or involve distinctions that we will never come across and think to factor into our deliberations. Furthermore, we shouldn't take it for granted that ordinary agents are always capable of recognising when two propositions are necessarily equivalent to one another, or when one proposition implies another. Plausibly, our credences and utilities are subject to the same kinds of representational limitations, and our limited rationality means that we're not always going to have the same degrees of confidence or desire towards equivalent propositions that we don't recognise as such. Nor, for that matter, should we think that credences are monotonic: if we cannot always recognise when P entails Q, then we cannot be expected to set our credences accordingly.

I suspect that most nowadays will share these basic assumptions. They immediately raise a number of metaphysical questions. What are preferences, credences, and utilities? Moreover, how do they relate to one another, and to other (mental and non-mental) phenomena? From a broadly functionalist approach to understanding the attitudes, providing answers for the latter class of questions is a crucial first step towards answering the former. It is with respect to this kind of question that representation theorems are poised to play a especially useful role.

Presumably, preferences and utilities are intimately tied to one another. Certain facts about utilities seem to necessitate facts about preferences. If S attaches a higher utility to P than she does to Q, then S prefers P to Q. And if S attaches the same utility to P and to Q, then she is indifferent between the two. These two claims seem like analytic truths, if any are.

Things aren't so obvious in the other direction. It looks reasonable to say that the *higher utility* relation is transitive and asymmetric, and that *equal utility* is an equivalence relation. But perhaps S strictly prefers P to Q, while the rest of her system of preferences is too ill-behaved to make sense of any kind of absolute utility assignment to either proposition—as would plausibly be the case if she preferred P to Q, Q to R, and R to P, being indifferent with respect to everything else. In short, it seems possible for S's preference relations to have properties which come apart from those we would expect of relative utility relations. Preferences can be intransitive; relative utilities cannot be.[4]

One plausible condition, then, for the sensible assignment of utilities is a *minimally well-behaved preference structure*: strict preferences ought to be transitive and asymmetric; indifferences ought to be transitive, symmetric and reflexive. Some preference structures are not so well-behaved. It seems, then, that we can have preferences in the absence of utilities, but we can't have an assignment of utilities in the absence of preferences. This asymmetry provides some *prima facie* support for the metaphysical thesis that facts about preferences are more fundamental than facts about utilities.

But we needn't posit anything quite so strong as this just yet. What's more important for our immediate purposes is that the same kinds of necessary connections between utilities and preferences just noted imply that information about a subject's preferences puts limiting constraints on the facts about her utilities. If S prefers P to Q, then she can't have less or equal utility for P than she does for Q—that is, if she has utilities at all. And if S is indifferent between P and Q, then she can't have a greater utility towards either one over the other. The natural question to ask at this point concerns just *how much* information about our utilities can be extracted from our preferences. If it turns out that complete knowledge of subjects' preferences fully determines the facts about their utilities, that would certainly suggest pursuing a functional definition of the latter in terms of the former. And even if that information *doesn't* pin down everything we might want to know, it may at least point in the direction of where to look for further constraints.

As it turns out, we can be pretty confident that there must be more to the story of having such-and-such utilities than just having a minimally well-behaved preference structure (in the sense of 'minimally well-behaved' just outlined). Utilities represent

[4]There is room for disagreement here. It's easier to argue that a subject's preferences can be very ill-behaved when these are thought of as representing choice dispositions. But things are not so straightforward when preferences are understood as mental states, for which we only have intuitive evidence to rely on. To be sure, it is certainly very hard to imagine a strict preference relation which is not asymmetric; likewise an indifference relation which is not symmetric. I'm inclined to think that these properties are constitutive of strict preference and indifference, respectively. But it is much more plausible that transitivity of preference can sometimes fail, and that is what I am mainly appealing to here. Where transitivity fails, one *might* argue that we can still make sense of *local* or *context-dependent* utilities, even though a global numerical representation of the subject's preferences won't be possible. I suspect that something like this is probably right, but it also very naturally fits the picture where preferences are prior to, and part of the grounds of, any correct assignment of utilities.

not just an order of preference, but also the varying strengths of those preferences. And without some further constraints, we won't yet get enough information out of the preferences needed to capture the *extra*-ordinal information that we ordinarily assume is encoded in a proper assignment of utilities. Ramsey [34, p. 176] recognised this problem nearly a century ago. He also proposed a solution, which has since become standard (see, e.g., [42]): given a plausible picture of how preferences are formed under conditions of uncertainty, it seems that extra-ordinal facts about our utilities are primarily functionally relevant in conditions of uncertainty. So, to get a fix on the extra-ordinal utility facts, we need a fix on at least some facts about our credences.

Remarkably, Ramsey also purported to show that the relevant facts about a subject's credences were *themselves* derivable from facts about that subject's preferences, at least under certain conditions and given some reasonable assumptions about how credences and utilities generally interact in the production of preferences. This was the intended upshot of his representation theorem, and just below I will discuss how this might work in some detail. Numerous authors have since established similar results; e.g., Savage [36], Anscombe and Aumann [2], Jeffrey [21], Luce and Krantz [28], to name just a small few. See also [16] for a helpful overview.

Now, as a matter of fact I don't think that the facts about our credences and utilities are *fully* grounded in the facts about our preferences. Partly, this is because I think that not one of the representation theorems we currently have actually supports the kind of derivation that Ramsey thought he could provide. But, moreover, we shouldn't think that credences are fully grounded in preferences because to do so would neglect the important role that credences have in response to perceptual evidence and *a priori* reasoning. (Cf. Lewis [26] on the need for both Rationalisation *and* Charity considerations in fixing upon a correct interpretation of a subject's mental states). The 'input' side of the functionalists' equation is something which is going to be missed by any characterisation of credences given wholly in terms of the downstream causal effects that they are supposed to have.

But it *is* valuable, and instructive, to see how far we can get in pinning down our credences and utilities with preferences alone. As I will argue (see Section 5 esp.), it does appear that there are some facts about our credences—e.g., what set of propositions we have some credences towards in the first place—that we cannot straightforwardly extract from our preferences. Knowing what we can and can't get from preferences gives some indication of what else may be needed, and where to look for it. And for this, we need a representation theorem.

2.1 Representation Theorems: What We Want Versus What We've Got

Earlier, I said that the standard expected utility representation theorem tells us that if a subject's preferences satisfy certain axioms, then she can be represented as maximising her expected utility under a particular set of credences and utilities—and that having those credences and utilities is the only way that she could be an expected utility maximiser. Before moving on, it will be worth making this more precise.

The following is a generic interpretation of a non-specific decision-theoretic representation theorem, with strong uniqueness conditions:

Rep. Theorem If S's preferences w.r.t. domain $\mathscr{D}$ conform to axioms $\mathscr{A}$, then there is (in effect) exactly one pair of functions Cr and U satisfying restrictions $< R_1, R_2, \ldots, R_n >$ such that S is representable as following decision rule $\mathscr{R}$ (w.r.t. $\mathscr{D}$) given credences Cr and utilities U.

In the paragraphs that follow, I'll break this down into its parts and discuss each one, before considering how such a theorem might be deployed.

To begin with, every representation theorem posits a specific *domain* over which a subject's preferences are to be defined (i.e., to which the axioms are applied). We will label this domain '$\mathscr{D}$'. In some cases, $\mathscr{D}$ is taken to be the set of all things towards which the subject *actually* has preferences; sometimes it's the set of all things towards which anyone at all *might* have preferences. Both of these treatments tend to go along with extremely strong demands on the subject's overall preference structure. In other cases, $\mathscr{D}$ is taken to be a proper subset of the things towards which a person does or can have preferences. Generally speaking, we can take $\mathscr{D}$ to be whatever we want it to be—so long as it satisfies whatever structural assumptions we make of it.

Next, for the purposes of proving any representation result and its corresponding uniqueness result, two (closely related) assumptions must be made about the manner in which we are going to represent our subjects. First, we need to specify a *rule* according to which a subject's credences and utilities (which we'll label 'Cr' and 'U' respectively) are supposed to interact in generating her preferences. For the theorems philosophers are most familiar with, this rule will be some form of expected utility maximisation. There are, however, also many representation theorems for all sorts of non-expected utility theories. For instance, Buchak's [5] theorem has us represent subjects as conforming to her risk-weighted expected utility rule, which takes as input Cr, U, and the subject's attitudes towards risk as represented by a function R. To accommodate the variation between theorems, let's simply use '$\mathscr{R}$' to represent the particular decision rule that we are assuming our subject ought to be represented as following when deciding upon her preferences.

Second, we also require a specification of the *admissible properties* of the functions Cr and U (and whatever other functions, such as Buchak's R, which may be involved in the representation). For instance, common basic properties include that U be bounded and real-valued, and that Cr be a probability function on some prespecified algebra. (More examples will follow just below). These are restrictions on the *shape* and *domains* of Cr and U. To keep things general, let's use '$< R_1, R_2, \ldots, R_n >$' as a place-holder for whatever properties Cr and U are assumed to satisfy in establishing the representation result.

Given this, it is also critically important to emphasise that those same assumptions about $< R_1, R_2, \ldots, R_n >$ and $\mathscr{R}$ are needed for the purposes of establishing the theorem's uniqueness condition. In general, a uniqueness condition only tells us that Cr and U are unique (to whatever extent they are unique), *relative* to some assumptions about the shape and domains of those functions and the particular rule by which they are to be combined. For discussion on the importance of this point, see [46].

And finally, following orthodoxy I will assume that utilities are measurable only on an interval scale; i.e., U ought to be unique only up to positive linear transformation. Any positive linear transformation of U can then be said to represent no more

meaningful information that U does itself, so we can say that U is *in effect* unique just in case it is unique up to positive linear transformation. (This makes the statement of the generic representation result significantly simpler).

With the more precise statement of a representation theorem thus given, we can then put it to work with the following (very schematic) chain of reasoning:

1 Under conditions $\mathscr{C}$, S's preferences w.r.t. $\mathscr{D}$ conform to axioms $\mathscr{A}$
2 *From 1 and Rep. Theorem*: Under conditions $\mathscr{C}$, there is (in effect) exactly one pair Cr, U satisfying $< R_1, R_2, \ldots, R_n >$ such that S is representable as following rule $\mathscr{R}$ (w.r.t. $\mathscr{D}$) given credences Cr and utilities U
3 Under conditions $\mathscr{C}$, S does in fact follow rule $\mathscr{R}$ (w.r.t. $\mathscr{D}$)
4 If S follows rule $\mathscr{R}$ (w.r.t. $\mathscr{D}$), then S can be represented as such
5 *From 2, 3, and 4*: Under conditions $\mathscr{C}$, S either has credences Cr and utilities U, or her credences and utilities do not satisfy $< R_1, R_2, \ldots, R_n >$
6 Under conditions $\mathscr{C}$, S's credences and utilities satisfy $< R_1, R_2, \ldots, R_n >$
7 *From 5 and 6*: Under conditions $\mathscr{C}$, S has credences Cr and utilities U
8 If S would have credences Cr and utilities U under conditions $\mathscr{C}$, then she actually has credences and utilities (approximated by) Cr and U

$\therefore$ *From 7 and 8*: S has credences and utilities (approximated by) Cr and U

For theorems with weaker uniqueness conditions, the conclusion of this argument would need to be weakened correspondingly. For example, where Jeffrey's theorem only pins down a set of pairs $< Cr, U >$ related by a fractional linear transformation, the conclusion should be that S's actual credences and utilities can be located somewhere in that set.[5] The introduction of the 'conditions $\mathscr{C}$' and the final argumentative step involving premise 8 allows for a limited degree of idealisation in getting S to satisfy the axioms. Essentially: idealisations are allowed, but only to the extent that they do not involve anything that would significantly alter S's credences and utilities. An example of something that could go into $\mathscr{C}$ that I'll mention later is the assumption that S is risk-neutral. Another would be the assumption that S is not under the influence of any intoxicating substances, nor facing any time constraints or abnormal internal or external stresses.

Now, I take it that premise 4 will be entirely uncontroversial, an instance of the plausibly *a priori* principle that *actuality implies representability*. The strength of the argument therefore rests on the empirical premises—1, 3, 6 and 8—which in

[5] It is sometimes said that where a representation theorem does not determine a unique Cr and U, we ought to take the entire *set* of admissible Cr and U functions as our representation of the subject's credences and utilities respectively. Setting aside the sometimes questionable motivations for going this route, note that what's really going on here is a re-interpretation of the original theorem—i.e., not as saying that S can be non-uniquely represented as an expected utility maximiser with such-and-such credences and utilities (each represented by a single real-valued function), but instead as saying that S can be *uniquely* represented as following a more complicated decision rule with such-and-such credences and utilities (represented by sets of real-valued functions). The more complicated decision rule may be something like: prefer P to Q just in case the Cr-weighted average utility of P is greater than Q for each/some admissible Cr-U pair.

turn depend on the specifics of $\mathscr{C}$, $\mathscr{A}$, $\mathscr{D}$, $\mathscr{R}$ and $< R_1, R_2, \ldots, R_n >$.[6] In the best case scenario, we'd have a theorem with reasonably strong/interesting uniqueness conditions such that:

(i) The conditions $\mathscr{C}$ are the *actual conditions* (or at least *ordinary conditions*)
(ii) All proper functioning subjects satisfy axioms $\mathscr{A}$ with respect to $\mathscr{D}$
(iii) All proper functioning subjects plausibly follow $\mathscr{R}$ with respect to $\mathscr{D}$
(iv) $< R_1, R_2, \ldots, R_n >$ are minimal, so plausibly satisfied by an adequate representation of any proper functioning subject's credences and utilities

We do not currently have a representation theorem with these properties, and in more than one respect current theorems are quite far from it. This is the central point of a recent paper by Meacham and Weisberg [31] (see also [11, 12]). But there are also no arguments in the literature to suggest we *cannot* develop such a theorem, and I have explained already why I think it would be valuable to do so.

It's unlikely that we'll get things right the first time around; the best I think we can hope for right now is to edge ever closer by the successive de-idealisation of existing theorems. A theorem with axioms that ordinary subjects approximately satisfy relative to some interesting domain under minimally idealised conditions, with an intuitively plausible decision rule $\mathscr{R}$ and relatively few constraints $< R_1, R_2, \ldots, R_n >$ on Cr and U, would be a good start. Even better if it has reasonably strong uniqueness conditions. The development of just such a theorem is my aim in Sections 3 and 4.

One more point in connection with desideratum (iv), before we move on to the theorem. It's worth noting that *some* of the properties specified in $< R_1, R_2, \ldots, R_n >$ may reflect nothing more than a non-essential choice of notational system.[7] For instance, we don't *have* to represent credences using functions that only take values between 0 and 1, as opposed to 0 and 100, say. So in a perfectly good sense, the requirement that Cr takes values within [0, 1] is really just an artifact of our measuring conventions (cf. [22], pp. 414ff). For want of a better term, call these *conventional properties*. There is nothing substantive at issue when we need to assume that Cr and U satisfy conventional properties when we're proving a representation result.

However, there will generally also be some non-conventional properties present in $< R_1, R_2, \ldots, R_n >$. Every existing representation theorem seems to require Cr and U to satisfy certain structural properties which are not clearly 'conventional'.

[6] I do not mean to imply that the deductive argument I have presented is the *only* way to put a representation theorem to work in fixing a subject's credences and utilities. For instance, one might try to approach the matter via inference to the best explanation. In the event that S satisfies (or comes close to satisfying) $\mathscr{A}$, perhaps the best explanation is that she follows $\mathscr{R}$ with credences Cr and utilities U. The deductive argument I've given here is meant to be illustrative, to help us draw out the kinds of properties a theorem should have if it is to be usefully applied in the relevant way. Even on the IBE model, we'll still want something *like* the desiderata (i) to (iv) I've outlined to hold—e.g., if $\mathscr{R}$ were relatively implausible, or $< R_1, R_2, \ldots, R_n >$ excessively strong, then we wouldn't have a very good explanation of S's preferences.

[7] A similar point holds of course for the particular way in which the decision rule $\mathscr{R}$ is formulated, which is naturally dependent on how Cr and U are characterised.

For example, because his particular system cannot accommodate gambles on non-contingent propositions, Ramsey has to *stipulate* *Cr*'s values for necessary and impossible propositions, and his uniqueness result only holds relative to that stipulation (see [34, p. 180], cf. [15, Section 2.6]). Jeffrey assumes that the domain of his *Cr* function is just the same as the domain of his *U*, which is hardly obvious and certainly not a mere notational matter. And in order to ensure his seemingly quite strong uniqueness result, Savage restricts the range of admissible *Cr* functions to probability functions in particular.[8]

A final example for the unconvinced. Savage also requires his *Cr* to be defined over an algebra of *events*—i.e., disjunctions of highly specific 'states' which must be (i) probabilistically independent of whichever act the subject might choose, and (ii) logically independent of the outcomes that might result. Furthermore, his *U* is to be defined over a set of *outcomes* which are supposed to be maximally specific with respect to what the subject cares about. Importantly, these states and outcomes are used to define the basic relata of Savage's preference relation (functions from states to outcomes). So, what kinds of propositions get counted as states and outcomes has to be built in to the interpretation of the theorem's basic formal structures from the outset, rather than derived from the subject's preferences. Consequently, in merely setting up his formal representation of a subject *S*'s preferences, Savage makes substantive assumptions about how *S* conceives of her choice situations—of what acts are available for choice, what states of affairs will be independent of the decision, and what kinds of things make a difference to what she cares about. As many have noted, these assumptions are not innocent (e.g., [17, 37]).

3 A Theorem for the Frequently Irrational

In what follows, I will show that if a system of preferences over a suitably characterised domain $\mathscr{D}$ (consisting of a set of *outcomes* and simple *gambles*) satisfies my axioms, then those preferences can be represented as maximising expected utility with respect to an effectively unique set of credences and utilities. In Section 3.1, I will begin with a brief discussion on the intended interpretation of the theorem's main formal elements. Then, in Section 3.2, I will outline three purely structural axioms, used to specify the set of gambles with which we will be interested. In Sections 3.3 and 3.4, I lay down the axioms needed for the construction of the utility function, and in Section 3.5 I add one more axiom needed for the construction of a credence function. Finally, in Section 3.6, I will complete the theorem with a final axiom, and discuss its various interpretations. Some modifications will be discussed in Section 4.

[8] See Meacham and Weisberg [31, pp. 657–659] for an argument that this restriction to probability functions in Savage is substantive, rather than merely notational. There are a number of issues here regarding what *exactly* Savage needed to assume about *Cr*, and what specific properties of his *Cr* might be conventional rather than substantive (e.g., whether additivity *per se* is conventional or not is controversial). I don't want to rest too heavily on this one example; the other examples should suffice to make the point.

3.1 Interpretational Preliminaries

For a given subject S, let $\succsim$ represent S's *weak preference* relation; $P \succsim Q$ if and only if S holds P to be at least as good as Q. We will use $\succ$ and $\sim$ for S's *strict preference* and *indifference* relations, respectively. (In Section 3.4, $\succ$ and $\sim$ will be defined in terms of $\succsim$, in the usual way). The domain of these preference relations is a space of *propositions*.

I want to remain as neutral as possible regarding different theories about the nature of propositions, by which I simply mean truth-evaluable *objects of thought*. I do not assume that they are sets of possible (and perhaps impossible) worlds, ordered n-tuples of properties and objects, structures of Fregean senses, or what have you. Most importantly, I am allowing—though not assuming—that there can be numerically distinct but necessarily equivalent propositions. Consequently, in what follows I will keep things neutral between a coarse-grained approach to propositions (according to which logically equivalent propositions are identical), and a fine-grained approach (where, e.g., P and $\neg\neg P$ might represent distinct objects of thought).

For reasons that will become clearer as we move on, I do make some minimal assumptions about propositions. I assume that for any proposition P, it makes sense to talk of its negation, $\neg P$. And I assume that for any pair of propositions P and Q, it makes sense to speak of their conjunction, $P \wedge Q$. I doubt that any adequate theory of propositions will deny me these assumptions. I also assume that it makes sense to embed certain pairs of propositions within counterfactual conditionals. But again, there is nothing very committal here.

Let $\mathcal{O} = \{o_1, o_2, o_3, \ldots\}$ be a set of propositions over which our subject S has preferences; this will form the domain of our U function. (The use of '$\mathcal{O}$' is to indicate that its elements will form the *outcomes* of gambles, as discussed shortly). With one exception to be discussed in Section 3.2, no special assumptions need to be made about $\mathcal{O}$'s internal logical structure. $\mathcal{O}$ does not have to be closed under entailment, negation, disjunction, and so on, nor do we have to assume that the propositions in $\mathcal{O}$ have very *specific* contents. For simplicity of exposition, I will adopt the notational convention that *sameness of subscript implies sameness of desirability* (but not vice versa). For instance, it should be assumed in all that follows that o_1' and o_1'' each refer to outcomes with the same desirability as o_1 (i.e. $o_1 \sim o_1' \sim o_1''$). It should *not* be assumed that o_1' and o_1'' are always distinct from o_1.

Let $\mathcal{P} = \{P, Q, R, \ldots\}$ be another set of propositions; this will form the domain of our Cr function (and contains the 'win conditions' in the gambles that I will describe shortly).[9] As with $\mathcal{O}$, the formal treatment of $\mathcal{P}$ is compatible with many views on the nature of propositions, so necessarily equivalent propositions may end up constituting distinct members of $\mathcal{P}$. We will, however, end up having to require that $\mathcal{P}$ is closed under negation (this is a consequence of Section 2.1 and A1.2). It

[9] Because we have to specify $\mathcal{P}$ at the outset, the following theorem cannot really be thought of as giving us a way of deriving a subject's credences from her preferences. Instead, we can say that *given* knowledge of what propositions S has some credences, the theorem allows us to work out just what degrees of confidence she assigns to each. See Section 5 for further discussion.

would be possible to suppose that every proposition that is in $\mathcal{O}$ is also in $\mathcal{P}$ (and vice versa), but this will not be presumed. In this respect the theorem that follows differs from Jeffrey's, where the domains of Cr and U (and $\succsim$) are presumed identical. It also differs from Savage's theorem and a vast number of similar theorems, where it's assumed that the domains of Cr and U are non-overlapping.

My strategy is heavily influenced by Ramsey [34], whose theorem involves the extraction of an agent's credences and utilities from her preferences over a set of gambles with maximally (or near-maximally) specific propositions as payoffs, under the assumption that she is a logically omniscient and deductively infallible expected utility maximiser (see [15, Section 2]). My space of gambles will in one sense be less restricted (in another sense, more). In general, a *two-outcome gamble* can be thought of as any *act* or *choice* such that, if made, o_1 would end up being the case were P the case, and o_2 would end up being the case otherwise. Let $[o_1, P; o_2]$ represent such a gamble. On first pass, if we say that $\mathcal{G}$ is a (still-to-be-specified) set of two-outcome gambles, then we can say that the relevant domain $\mathscr{D}$ of $\succsim$ to which the axioms that follow will be applied is $\mathcal{O} \cup \mathcal{G}$. Exactly which gambles get into $\mathcal{G}$ will be discussed in Sections 3.2 and 3.6.

We need to be a little careful here, for two reasons. First of all, it's possible to be mistaken about a gamble's payoff structure—about the pattern of outcomes that would result should it be chosen—and so we shouldn't characterise gambles according to payoff structure that they *actually* have. We value gambles according to the payoffs we *believe* they'll have, not according to the payoffs they actually have. Secondly, and more importantly, I have characterised $\succsim$ as a relation between propositions—and gambles are not propositions. Thus, it's not quite correct to say that $\mathcal{G}$ is a set of gambles. Instead, I will assume that the value S attaches to a gamble $[o_1, P; o_2]$ with a *known* payoff structure is the same value she attaches to the state of affairs *that she has made the gamble* $[o_1, P; o_2]$. Strictly speaking, $\mathcal{G}$ only includes the latter. In practice, I'll continue to speak as though $\mathcal{G}$ is a set of gambles, rather than propositions about S having made such-and-such a gamble. Likewise, I'll use $[o_1, P; o_2]$ (and etc.) to represent members of $\mathcal{G}$ directly.

The central goal will be to find a set of axioms for $\succsim$ on $\mathcal{O} \cup \mathcal{G}$, such that S's value for $[o_1, P; o_2]$ is determined by the utility that she attaches to the outcomes (under the relevant conditions) weighted by the uncertainty she has for the gamble's conditions:

$$U([o_1, P; o_2]) = Cr(P) \cdot U(o_1 \wedge P) + Cr(\neg P) \cdot U(o_2 \wedge \neg P)$$

In supposing that S's utilites are determined in this way, I am assuming that S is neither risk seeking nor risk averse (when considering the relevant gambles). To the extent that this is false, we really ought to be considering S's preferences *were* she to have risk neutral attitudes (this would go into the specification of the conditions $\mathscr{C}$).

3.2 Purely Structural Axioms

The way in which we formalise the members of $\mathcal{G}$ is not especially important, but it will be helpful to think of them as functions from complementary pairs of propositions in $\mathcal{P}$ to outcomes in $\mathcal{O}$. So, $[o_1, P; o_2]$ is just the set $\{(P, o_1), (\neg P, o_2)\}$, and

consequently, $[o_1, P; o_2]$ and $[o_2, \neg P; o_1]$ will always be equipreferable, being just two different ways of representing one and the same gamble.

What's more important is that $\mathcal{G}$ as a whole be *suitably characterised*. The main purpose of this section is to explain what I mean by that, by means of three background *purely structural* axioms. These are axioms which impose no direct constraints on $\succsim$, instead serving to characterise some basic properties which we'll be assuming hold of $\mathcal{O}$, $\mathcal{P}$ and $\mathcal{G}$. There are also several *partially structural* axioms which will follow later; see A1, A5, and A8.

To begin with, although I have placed very few conditions on the *internal* structures of $\mathcal{O}$ and $\mathcal{P}$, it is important for the results that follow that the propositions in $\mathcal{O}$ might stand in a certain kind of *relation* to the propositions in $\mathcal{P}$. We'll denote this relation using '$\rightharpoonup$', and in the event that $P \rightharpoonup Q$ and $Q \rightharpoonup P$, we will write '$P \rightleftharpoons Q$'. I will discuss the interpretation of '$\rightharpoonup$' in some depth in Section 3.6—there are several interesting possibilities here, and each gives rise to a different way of understanding the theorem as a whole. For now, it will be most helpful to read '$P \rightharpoonup Q$' under any of the following interpretations, listed in order of strength:

Interpretation 1 S believes that P implies Q
Interpretation 2 S recognises that P implies Q
Interpretation 3 P obviously implies Q

By 'P obviously implies Q', I mean that we can reasonably expect any ordinary subject capable of considering the matter to *recognise* that P implies Q, where *recognition* is a factive species of belief. For example, we can reasonably expect of anyone capable of contemplating that P to know that P implies that $P \wedge P$, but it may not be so obvious that P implies $P \rightarrow \neg((\neg P \wedge Q) \wedge (P \wedge \neg Q))$. Under none of the suggested interpretations should we presume that $\rightharpoonup$ is transitive. In Section 3.6, I'll suggest some reasons to prefer Interpretation 1, though I think each has interest.

In the very final stages of the theorem's proof, I will make appeal to three presumed characteristics of $\rightharpoonup$. These I state now as a single structural assumption:

Axiom S1 For all $P, Q \in \mathcal{P} \cup \mathcal{O}$,

(S1.1) If $P \rightharpoonup Q$, then $P \rightleftharpoons (P \wedge Q)$
(S1.2) If $P \rightleftharpoons Q$, then $\neg P \rightleftharpoons \neg Q$
(S1.3) If $P \rightleftharpoons Q$, then if $R \rightharpoonup P$, $R \rightharpoonup Q$

I suspect that there will be counterexamples to S1 under any of the three interpretations I've suggested. The most problematic clause is S1.3, which by itself implies that $\rightleftharpoons$ is transitive. Furthermore, if we also assume that $P \rightharpoonup Q$ whenever $P \rightharpoonup (Q \wedge R)$, which seems reasonable enough, then S1.1 and S1.3 together imply that $\rightharpoonup$ is transitive. However, the latter two clauses only play a very minor role in what follows, and the nature of that role is such that it's more important that they come *close* to the truth than that they hold with absolutely no exceptions. In Section 3.6, I'll briefly note how S1.2 and S1.3 might be dropped in favour of something less problematic. S1.1, on the other hand, is crucial, for reasons that I will get to shortly.

Once we've pinned down $\rightharpoonup$, the next structural assumption directly characterises the kinds of gambles which can be found in $\mathcal{G}$:

Axiom S2 $[o_1, P; o_2] \in \mathcal{G}$ iff:

(S2.1) $o_1, o_2 \in \mathcal{O}$ and $P, \neg P \in \mathcal{P}$
(S2.2) $o_1 \rightharpoonup P$ and $o_2 \rightharpoonup \neg P$
(S2.3) If P is possible, so is $o_1 \wedge P$, and if $\neg P$ is possible, so is $o_2 \wedge \neg P$

Axiom S2 does not by itself imply that $\mathcal{G}$ contains any gambles; it is merely meant to characterise the *kinds* of gambles that *might* go into $\mathcal{G}$. Later, A1 will state that for every P in $\mathcal{P}$, $\mathcal{G}$ contains a gamble on P; and for every outcome o, there will be a gamble in $\mathcal{G}$ with an outcome *equal in value* to o.

S2.1 is not particularly interesting, requiring only that the gambles in $\mathcal{G}$ be constructed from the members of $\mathcal{O}$ and $\mathcal{P}$. S2.2 and S2.3, on the other hand, are somewhat more substantive. S2.3 has us limiting our attention to those gambles $[o_1, P; o_2]$ where the outcomes o_1 and o_2 are $\rightharpoonup$-related to the conditions under which they are won. Given S1.1, this means that $o_1 \rightleftharpoons (o_1 \wedge P)$ and $o_2 \rightleftharpoons (o_2 \wedge \neg P)$, which will ultimately be used to ensure that if $[o_1, P; o_2] \in \mathcal{G}$, then $U(o_1) = U(o_1 \wedge P)$ and $U(o_2) = U(o_2 \wedge \neg P)$. Indeed, more generally I will eventually want to establish that (where P and Q are assigned credences and utilities at all), if $P \rightleftharpoons Q$, then $Cr(P) = Cr(Q)$ and $U(P) = U(Q)$. Implicit in this is a centrally important constraint on our interpretation of $\rightharpoonup$; *viz.*, it needs to be a relation such that, plausibly, the above equalities hold. Where '$\rightharpoonup$' is take to represent *obvious*, *recognised*, or even just *believed* implication, this seems fair: one who thinks that P and Q imply each other should assign the same credences and utilities to P and Q. A failure to treat them as such suggests a lack of belief in their equivalence.[10]

The purpose of S2.3 is to rule out what we might call *impossible gambles*. Not every 'gamble' that we might be able to formally construct out of the members of

[10]There may be some difficulties here regarding *framing effects*, whereby a choice might be evaluated differently depending on whether its outcomes are cast in a negative or a positive light (see [23, 40]). For example, a doctor might know that giving a population of 1000 deathly ill patients a particular treatment will cure 75 % but kill the rest. When choosing whether to administer the treatment, it seems to make a difference whether this outcome is described as '750 lives are saved' or as '250 people die', although in both cases the doctor presumably recognises that 750 will live and 250 will die. We do not know the mechanisms underlying these effects, so it's unclear whether they conflict with the assumption that $U(P) = U(Q)$ whenever $P \rightleftharpoons Q$. One plausible explanation which doesn't obviously generate conflict is that the way in which a choice is framed can make particular *aspects* of a complex outcome more salient than other aspects [25, 41]. So, instead of representing the doctor as assigning different utilities to distinct but recognisably equivalent representations of one and the same outcome (*750 will live & 250 will die*), we see her as having different utilities towards non-equivalent aspects of the outcome (*750 will live*, *250 will die*), with positive or negative descriptions of that outcome influencing which aspects get represented as 'the' outcome. If this kind of explanation is correct, then framing effects describe an error in how agents go from *descriptions* of choices to their own internal *representations* of those choices. Since my $\succsim$ is defined over the representations directly, we do not have to worry about any potential cognitive biases that might influence how we go from a description of a gamble or outcome to the (mis-)representation thereof.

$\mathcal{O}$ and $\mathcal{P}$ represents a genuine object of choice. When a decision-maker takes up a gamble $[o_1, P; o_2]$, she is making true the following conjunction of counterfactuals:

$$(P \mathbin{\square\!\!\rightarrow} o_1) \wedge (\neg P \mathbin{\square\!\!\rightarrow} o_2)$$

However, not every conjunction of this form is possibly true. Assuming a Lewisean semantics for counterfactuals (with a space of epistemically possible worlds), S2.3 ensures that every gamble in $\mathcal{G}$ corresponds to an epistemically possible conjunction of counterfactuals.[11] We can call a gamble $[o_1, P; o_2]$ *impossible* just in case P is possible and $o_1 \wedge P$ isn't, and/or $\neg P$ is possible and $o_2 \wedge \neg P$ isn't; it's *possible* otherwise. Note that S2.3 does *not* rule out gambles with impossible win conditions, nor does it rule out gambles with impossible outcomes—*possible* gambles may have *impossible* parts!

Finally, we will need one last purely structural axiom. I have said just above that I intend to show that if $[o_1, P; o_2] \in \mathcal{G}$, then $U(o_1) = U(o_1 \wedge P)$ and $U(o_2) = U(o_2 \wedge \neg P)$. This can only be true if $U(o_1 \wedge P)$ and $U(o_2 \wedge \neg P)$ are defined, which will require the relevant propositions to be in $\mathcal{O}$:

Axiom S3 If $[o_1, P; o_2] \in \mathcal{G}$, then $o_1 \wedge P \in \mathcal{O}$

On the fairly plausible presumption that $(P \wedge Q) \nsim P$, S2 and S3 jointly imply that $\mathcal{O}$ contains infinitely many fine-grained propositions: for any $[o_1, P; o_2]$ in $\mathcal{G}$ where P is consistent, there will be $o_1 \wedge P$ and $o_2 \wedge \neg P$ in $\mathcal{O}$ that (are believed to) imply P and $\neg P$ respectively. It is then easy to check that S2.1, S2.3, and S2.3 will also hold for the gamble $[o_1 \wedge P, P; o_2 \wedge \neg P]$, so S3 then requires that $\mathcal{O}$ also include $(o_1 \wedge P) \wedge P$ and $(o_2 \wedge \neg P) \wedge \neg P$, and so on, *ad infinitum*.

I don't think that this should be cause for much concern. For one thing, it's not obvious that $o_1 \wedge P$, $(o_1 \wedge P) \wedge P$, and so on, constitute genuinely distinct *objects of thought*. But even if you take a very fine-grained approach to content, it's not clear there is a real worry here. Although before now I have never considered the propositions before, in a *dispositional* or *implicit* sense I have very plausibly always believed that *1523410 is one less than 1523411*, and that *1523411 is one less than 1523412*, and so on. Likewise, I have always known that $P \wedge Q$ is just the same state of affairs as $(P \wedge Q) \wedge Q$, and $((P \wedge Q) \wedge Q) \wedge Q$, and so on, *ad infinitum*. If so, then it's not implausible that a function U designed to represent my utilities should assign one and the same value to an infinite collection of propositions that only differ by successive conjunctions of the same conjunct—even if, for the vast majority of these propositions, I have not and likely never will take the time to contemplate them explicitly.[12]

[11] I do not place very much weight on this assumption about the semantics of counterfactuals. For instance, if there can be impossible counterfactuals with impossible antecedents, then alternative conditions can be placed on $\mathcal{G}$ to fix upon the appropriate set.

[12] A similar point holds, I think, for the equivalence between P, $\neg\neg P$, $\neg\neg\neg\neg P$ (and etc.). To the extent that these represent distinct objects of thought, it's reasonable to think that most ordinary agents know (at least implicitly) that if the number of negations preceding a claim P is a multiple of two, then the proposition expressed is equivalent to P; otherwise it's equivalent to $\neg P$.

3.3 $\frac{1}{2}$ Probability Propositions

For the next step, we will need to characterise a set of propositions towards which our subject has a credence of $\frac{1}{2}$, which we will label using 'Π'. To do this "without peeking", we'll need a way to construct Π given just S's preferences and the minimal assumptions we've made about her. The strategy we'll use was pioneered by Ramsey [34, pp. 177–8], though I'll appeal to Savage's generalisation of the same reasoning.

Suppose we have two outcomes o_1 and o_2 such that $o_1 \succ o_2$, and two gambles $[o_1, P; o_2]$ and $[o_1', Q; o_2']$ (both in $\mathcal{G}$), with $[o_1, P; o_2] \succsim [o_1', Q; o_2']$. (Recall that sameness of subscript implies indifference, so $o_1' \sim o_1 \succ o_2' \sim o_2$). On the assumptions we've made so far, if our subject is maximising expected utility with respect to these two gambles, then she must attach at least as much credence to P as she does to Q. Essentially, $[o_1, P; o_2]$ is weakly preferred to $[o_1', Q; o_2']$ just because it has at least as great a chance of resulting in the better outcome o_1 as $[o_1', Q; o_2']$ does.

That gives us the general case, and we will eventually be able to show that $Cr(P) \geq Cr(Q)$ will hold whenever there are outcomes and gambles such that $o_1 \succ o_2$ and $[o_1, P; o_2] \succsim [o_1', Q; o_2']$.[13] For now, take the special case where $Q = \neg P$, and suppose that $[o_1, P; o_2] \sim [o_1', \neg P; o_2']$. The latter of these two gambles is just another way of representing $[o_2', P; o_1']$, so $[o_1, P; o_2] \sim [o_2', P; o_1']$. Since we know that $o_1 \succ o_2$, the only way to get this kind of indifference between the gambles is if $Cr(P) = Cr(\neg P) = \frac{1}{2}$. Hence:

Definition 1 $\Pi = \{P \in \mathcal{P}$: for some $[o_1, P; o_2], [o_2', P; o_1'] \in \mathcal{G}$ such that $o_1 \nsim o_2$, $[o_1, P; o_2] \sim [o_2', P; o_1']\}$

Henceforth, I will use π, π', π'', and so on, to designate propositions within Π. (It should not be assumed that $\pi \neq \pi'$).

Many of the axioms I'll outline relate specifically to gambles conditional on a $\frac{1}{2}$-probability proposition, and so it is important to ensure that Π contains enough members. We do this by means of the following axiom:

Axiom A1 $\mathcal{O}$ and $\mathcal{P}$ are non-empty, and:

(A1.1) For every pair $o_1, o_2 \in \mathcal{O}$, there is some $[o_1', \pi; o_2'] \in \mathcal{G}$
(A1.2) For all $P \in \mathcal{P}$, there is at least one $[o_1, P; o_2] \in \mathcal{G}$, where $o_1 \nsim o_2$

A1.1 implies that Π is non-empty, and that not all outcomes are equipreferable. Moreover, A1.1 is that each of the universally quantified statements about possible gambles conditional on some π in the proofs that follow are never trivially true. Given this, I will for ease of exposition often omit steps involving A1.1 when obvious. A1.2 will not be relevant until Section 3.5.

[13] To ensure that $Cr(P) \geq Cr(Q)$, it suffices to assume that $o_3 \succsim o_1 \succ o_4 \succsim o_2$ and $[o_1, P; o_2] \succsim [o_3, Q; o_4]$. Letting o_3 and o_4 be o_1' and o_2' respectively makes the reasoning somewhat more transparent, especially when it comes to defining Π.

One of A1.1's direct consequences is that for each outcome o_1, there will be at least two outcomes o_1' and o_1'' such that $o_1' \rightharpoonup \pi$ and $o_1'' \rightharpoonup \neg\pi$. It's plausible that for a great many values we will be able to find such a proposition. Consider, for instance, the following situation. Our subject has no intrinsic interest in the outcomes of coin tosses. Let o be an arbitrary consistent outcome; and let π be the proposition *the next fair coin to be tossed lands heads*. Then, suppose that $o_1' = o_1 \wedge \pi$, while $o_1'' = o_1 \wedge \neg\pi$. Plausibly, $o_1 \sim o_1' \sim o_1''$, while o_1' (obviously) implies π and o_1'' (obviously) implies $\neg\pi$, but neither π nor $\neg\pi$ imply either o_1' or o_1''. The condition seems to be at least *approximately* satisfied in this sense—for any outcome o_1, we should be able to find a pair of equally-valued outcomes which are equivalent in all respects that the agent cares about but for the outcome of a fair coin toss (or some other $\frac{1}{2}$ probability event that our subject doesn't really care about).

3.4 Difference Relations and Algebraic Difference Structures

For the next stage, we will need to supply a condition for when the difference in utility between two outcomes o_1 and o_2 is at least as great as the difference in utility between another pair of outcomes o_3 and o_4. We will write this as $(o_1, o_2) \geq^d (o_3, o_4)$. As with our definition of Π, the trick is to assume the final form of the representation we want to end up with, and work backwards from that to a definition of $\geq^d$. That is, we suppose first of all that our subject is an expected utility maximiser with respect to the relevant gambles, so:

$$U([o_1, \pi; o_2]) = Cr(\pi) \cdot U(o_1 \wedge \pi) + (1 - Cr(\pi)) \cdot U(o_2 \wedge \neg\pi)$$

We then know that $[o_1, \pi; o_4] \gtrsim [o_2, \pi; o_3]$ just in case:

$$\begin{aligned} &Cr(P) \cdot U(o_1 \wedge \pi) + (1 - Cr(\pi)) \cdot U(o_4 \wedge \neg\pi) \geq \\ &Cr(\pi) \cdot U(o_2 \wedge \neg\pi) + (1 - Cr(\pi)) \cdot U(o_3 \wedge \neg\pi) \end{aligned}$$

Since $Cr(\pi) = \frac{1}{2} = 1 - Cr(\pi)$, and since $U(o_1 \wedge \pi) = U(o_1)$ (and so on) for the gambles that we will be considering, this holds just in case:

$$U(o_1) - U(o_2) \geq U(o_3) - U(o_4)$$

This gives us enough to characterise $\geq^d$ (the coherence of which will be established shortly below):

Definition 2 $[\geq^d]$ $(o_1, o_2) \geq^d (o_3, o_4)$ iff, for all $[o_1', \pi; o_4']$, $[o_2', \pi'; o_3'] \in \mathcal{G}$, $[o_1', \pi; o_4'] \gtrsim [o_2', \pi'; o_3']$

With the information about differences in utilities thus codified, we can look for ways to construct U. We will want the following result, established in [24, Section 4.4.1]:

Theorem 1 *Suppose that* $< \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}, \gtrsim^* >$ *is an algebraic difference structure; i.e.,* $\boldsymbol{\mathcal{X}}$ *is non-empty,* $\gtrsim^*$ *is a binary relation on* $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$*, and the following five conditions hold for all* $x_1, x_2, x_3, x_4 \in \boldsymbol{\mathcal{X}}$*, and all sequences* $x_1, x_2, \ldots, x_i, \ldots \in \boldsymbol{\mathcal{X}}$*,*

C1 $\succsim^*$ *on* $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$ *is transitive and complete*
C2 *If* $(x_1, x_2) \succsim^* (x_3, x_4)$*, then* $(x_4, x_3) \succsim^* (x_2, x_1)$
C3 *If* $(x_1, x_2) \succsim^* (x_4, x_5)$ *and* $(x_2, x_3) \succsim^* (x_5, x_6)$*, then* $(x_1, x_3) \succsim^* (x_4, x_6)$
C4 *If* $(x_1, x_2) \succsim^* (x_3, x_4) \succsim^* (x_1, x_1)$*, then there exist* $x_5, x_6 \in \boldsymbol{\mathcal{X}}$ *such that* $(x_1, x_5) \sim^* (x_3, x_4) \sim^* (x_6, x_2)$
C5 *If* $x_1, x_2, \ldots, x_i, \ldots$ *is such that* $(x_{i+1}, x_i) \sim^* (x_2, x_1)$ *for every* x_i, x_{i+1} *in the sequence,* $(x_2, x_1) \nsim^* (x_1, x_1)$*, and there exist* $x_j, x_k \in \boldsymbol{\mathcal{X}}$ *such that* $(x_j, x_k) \succ^* (x_i, x_1) \succ^* (x_k, x_j)$ *for all* x_i *in the sequence, then it is finite*

Then, there exists a function f: $\boldsymbol{\mathcal{X}} \mapsto \mathbb{R}$ *such that, for all* $x_1, x_2, x_3, x_4 \in \boldsymbol{\mathcal{X}}$*,*

(i) $(x_1, x_2) \succsim^* (x_3, x_4)$ *iff* $f(x_1) - f(x_2) \geq f(x_3) - f(x_4)$

Furthermore, f is unique up to positive linear transformation.

The main purpose of the next five axioms is to establish that $< \boldsymbol{\mathcal{O}} \times \boldsymbol{\mathcal{O}}, \geq^d >$ satisfies C1 through to C5. A2 says that $< \boldsymbol{\mathcal{O}} \cup \boldsymbol{\mathcal{G}}, \succsim >$ is a weak order:

Axiom A2 For all $x, y, z \in \boldsymbol{\mathcal{O}} \cup \boldsymbol{\mathcal{G}}$,

(A2.1) If $x \succsim y$ and $y \succsim z$, then $x \succsim z$
(A2.2) Either $x \succsim y$ or $y \succsim x$

We can now define $\succ$ and $\sim$. Say that $x \succ y$ iff $x \succsim y$ and $\neg(y \succsim x)$; and $x \sim y$ iff $x \succsim y$ and $y \succsim x$. Axiom A2 then ensures that $\sim$ is an equivalence relation.

The completeness requirement (A2.2) is quite strong. It's *prima facie* implausible that many ordinary agents have complete preference rankings, even where the domain of that ranking is restricted just to those propositions that they have utilities towards. We will return to this briefly in Section 4.2, when we discuss the need to accommodate the apparent imprecision in our credences and utilities.

The next says that if we have a gamble $[o_1, \pi; o_2]$, then we're allowed to substitute either outcome for another of equal value, or one proposition π for another π', so long as the gamble that results is in $\boldsymbol{\mathcal{G}}$:

Axiom A3 If $[o_1, \pi; o_2], [o_2', \pi'; o_1'] \in \boldsymbol{\mathcal{G}}$, then $[o_1, \pi; o_2] \sim [o_2', \pi'; o_1']$

Axiom A3 states that we're allowed to change around the order of equally-valued outcomes within gambles conditional on a $\frac{1}{2}$-probability proposition without changing the gamble's value. Given A1.1 and A2, the axiom also immediately implies that any pair of gambles $[o_1, \pi; o_2]$ and $[o_1', \pi'; o_2']$ are always equipreferable (since there will be some $[o_2'', \pi''; o_1'']$ to which each is equipreferable).

Although nearly opaque in its present formulation, in light of the earlier axioms the next axiom more or less directly implies that $\geq^d$ is transitive (see proof below):

Axiom A4 If, for all $[o_1, \pi; o_4], [o_2, \pi'; o_3]$, (i) $[o_3, \pi''; o_6], [o_4, \pi'''; o_5] \in \boldsymbol{\mathcal{G}}$, $[o_1, \pi; o_4] \succsim [o_2, \pi\prime; o_3]$, and (ii) $[o_3, \pi''; o_6] \succsim [o_4, \pi'''; o_5]$, then, for all $[o_1', \pi^*; o_6'], [o_2', \pi^+; o_5'] \in \boldsymbol{\mathcal{G}}$, $[o_1', \pi^*; o_6'] \succsim [o_2', \pi^+; o_5']$.

With these three new axioms, we have enough already to show that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies C1 through to C3, above. The following lemma will prove helpful, and relies solely on A1 to A3:

Lemma 1 *If* $[o'_1, \pi; o'_4] \succsim [o'_2, \pi'; o'_3]$ *for any pair* $[o'_1, \pi; o'_4], [o'_2, \pi'; o'_3] \in \mathcal{G}$, *then* $(o_1, o_2) \geq^d (o_3, o_4)$

Proof Suppose that $[o'_1, \pi; o'_4] \succsim [o'_2, \pi'; o'_3]$ for some such pair in $\mathcal{G}$. From A3, for any $[o''_1, \pi''; o''_4], [o''_2, \pi'''; o''_3] \in \mathcal{G}$, then, $[o'_1, \pi; o'_4] \sim [o''_1, \pi''; o''_4]$ and $[o'_2, \pi'; o'_3] \sim [o''_2, \pi'''; o''_3]$. Thus, for all such pairs, $[o''_1, \pi''; o''_4] \succsim [o''_2, \pi'''; o''_3]$, which is just the right hand side of Definition 2. □

We can now prove that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies C1:

Proof For any pair of pairs $(o_1, o_4), (o_2, o_3)$, there will be $[o'_1, \pi; o'_4]$, $[o'_2, \pi'; o'_3]$, and since $\succsim$ is complete, either $[o'_1, \pi; o'_4] \succsim [o'_2, \pi'; o'_3]$ or $[o'_2, \pi'; o'_3] \succsim [o'_1, \pi; o'_4]$. From Lemma 1, if the former then $(o_1, o_2) \geq^d (o_3, o_4)$, and if the latter then $(o_3, o_4) \geq^d (o_1, o_2)$. So $\geq^d$ is complete. Next, suppose that $(o_1, o_2) \geq^d (o_3, o_4)$ and $(o_3, o_4) \geq^d (o_5, o_6)$. From Definition 2, this implies (for all relevant gambles) that $[o'_1, \pi; o'_4] \succsim [o'_2, \pi'; o'_3]$ and $[o'_3, \pi^*; o'_6] \succsim [o'_4, \pi^+; o'_5]$. For any pair of gambles $[o''_1, \pi''; o''_6], [o''_2, \pi'''; o''_5]$, A4 then requires that $[o''_1, \pi''; o''_6] \succsim [o''_2, \pi'''; o''_5]$, so $(o_1, o_2) \geq^d (o_5, o_6)$. So $\geq^d$ is transitive. □

Furthermore, we can also prove that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies conditions C2 and C3:

Proof Suppose $(o_1, o_2) \geq^d (o_3, o_4)$, so $[o'_1, \pi; o'_4] \succsim [o'_2, \pi'; o'_3]$. A1 ensures some $[o''_4, \pi^*; o''_1], [o''_3, \pi^+; o''_2]$ exist, and by A3, $[o''_4, \pi^*; o''_1] \sim [o'_1, \pi; o'_4]$ and $[o''_3, \pi^+; o''_2] \sim [o'_2, \pi'; o'_3]$. Substituting for equally valued gambles, $[o''_4, \pi^*; o''_1] \succsim [o''_3, \pi^+; o''_2]$, which implies $(o_4, o_3) \geq^d (o_2, o_1)$. So $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies C2. By similar reasoning, if $(o_1, o_2) \geq^d (o_3, o_4)$ then $[o'_1, \pi; o'_4] \succsim [o''_3, \pi^+; o''_2]$, so $(o_1, o_2) \geq^d (o_3, o_4)$ also implies that $(o_1, o_3) \geq^d (o_2, o_4)$. Supposing next that $(o_1, o_2) \geq^d (o_4, o_5)$ and $(o_2, o_3) \geq^d (o_5, o_6)$, it follows that $(o_1, o_4) \geq^d (o_2, o_5)$ and $(o_2, o_5) \geq^d (o_3, o_6)$. As $\geq^d$ is transitive, $(o_1, o_4) \geq^d (o_3, o_6)$. Thus, if $(o_1, o_2) \geq^d (o_4, o_5)$ and $(o_2, o_3) \geq^d (o_5, o_6)$, then $(o_1, o_4) \geq^d (o_3, o_6)$, and C3 is satisfied. □

We will need two further axioms to show that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies the structural axioms C4 and C5. A5 corresponds directly to C4:

Axiom A5 If there are $[o_3, \pi; o_1], [o'_1, \pi'; o_4], [o_2, \pi'', o'_3] \in \mathcal{G}$ such that $[o_3, \pi; o_1] \succsim [o'_1, \pi'; o_4] \succsim [o_2, \pi'', o'_3]$, then there are $[o_5, \pi^*; o''_3], [o'_4, \pi^+; o_6] \in \mathcal{G}$ such that $[o_5, \pi^*; o''_3] \sim [o'_1, \pi'; o_4]$ and $[o'_4, \pi^+; o_6] \sim [o_2, \pi'', o'_3]$

With this, $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies C4:

Proof Suppose (a) $(o_1, o_2) \geq^d (o_3, o_4)$ and (b) $(o_3, o_4) \geq^d (o_1, o_1)$, for any $o_1, o_2, o_3, o_4 \in \mathcal{O}$. (a) will hold only if, for all relevant gambles in $\mathcal{G}$, $[o'_3, \pi; o'_1] \succsim [o''_1, \pi'; o'_4]$; likewise (b) will hold only if for the relevant gambles $[o'_4, \neg\pi'; o''_1] \succsim [o_2, \pi'', o'_3]$. So, since $[o'_4, \neg\pi'; o''_1] \sim [o''_1, \pi'; o'_4]$, if $(o_1, o_2) \geq^d (o_3, o_4) \geq^d (o_1, o_1)$, then there will be the relevant gambles in $\mathcal{G}$ such that $[o'_3, \pi; o'_1] \succsim [o''_1, \pi'; o'_4] \succsim [o_2, \pi'', o'_3]$. A5 then asserts the existence of some $o_5, o_6 \in \mathcal{O}$ with corresponding gambles in $\mathcal{G}$ such that, by Lemma 1, $(o_1, o_5) =^d (o_3, o_4)$ and $(o_2, o_4) =^d (o_6, o_3)$. We've already seen that $(o_1, o_2) \geq^d (o_3, o_4)$ implies $(o_1, o_3) \geq^d (o_2, o_4)$, so from $(o_2, o_4) =^d (o_6, o_3)$ we get $(o_2, o_6) =^d (o_4, o_3)$. And we've already shown that C2 holds, so this gives us $(o_3, o_4) =^d (o_6, o_2)$. So, if $(o_1, o_2) \geq^d (o_3, o_4) \geq^d (o_1, o_1)$, then there must be some $o_5, o_6 \in \mathcal{O}$ such that $(o_1, o_5) =^d (o_3, o_4) =^d (o_6, o_2)$. □

In terms of the final representation, A5 says that if there are two outcomes o_1 and o_2 such that $U(o_1) - U(o_2) = n \geq 0$, then for *every* outcome o_3 there must be another outcome o_4 such that either $U(o_3) - U(o_4) = n$ or $U(o_4) - U(o_3) = n$. This is weaker than the assumption that Ramsey used to get C4, which in our system would be:

Axiom A5* For every triple $o_1, o_2, o_3 \in \mathcal{O}$, there is an $o_4 \in \mathcal{O}$ such that for some $[o'_1, \pi; o'_3], [o_4, \pi'; o'_2] \in \mathcal{G}$, $[o'_1, \pi; o'_3] \sim [o_4, \pi'; o'_2]$

Roughly, A5* says that if you've got two outcomes o_1 and o_2 with $U(o_1) - U(o_2) = n$, then, for any other outcome o_3, you should be able to find an o_4 such that $U(o_3) - U(o_4) = n$. This has the unfortunate consequence of precluding the existence of highest and lowest ranked outcomes, whereas our A5 does not.

Finally, we will need to make sure that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies the C5. In effect, this condition ensures that the numerical representation U satisfies the Archimedean property of the reals: for any positive number x, and any number y, there is an integer n such that $n + x \geq y$. To cash out the next axiom, I will need to characterise a *strictly bounded standard sequence*:

Definition 3 (**Strictly bounded standard sequence**) $o_1, o_2, \ldots, o_i, \ldots$ is a strictly bounded standard sequence iff:

(i) For all $[o'_2, \pi; o'_1], [o''_1, \pi'; o'''_1] \in \mathcal{G}$, $[o'_2, \pi; o'_1] \nsim [o''_1, \pi'; o'''_1]$; and
(ii) For every o_i, o_{i+1} in the sequence, $[o'_{i+1}, \pi; o'_2] \sim [o'_1, \pi'; o'_i]$ for all $[o'_{i+1}, \pi; o'_2], [o'_1, \pi'; o'_i] \in \mathcal{G}$; and
(iii) There exists $o_j, o_k \in \mathcal{O}$ such that for all o_i in the sequence, $[o'_j, \pi; o'_i] \succ [o'_1, \pi'; o'_k]$ and $[o''_i, \pi''; o''_k] \succ [o''_j, \pi'''; o''_1]$, for any $[o'_j, \pi; o'_i], [o'_1, \pi'; o'_k], [o''_i, \pi''; o''_k], [o''_j, \pi'''; o''_1] \in \mathcal{G}$

The next axiom is then easy to state:

Axiom A6 All strictly bounded standard sequences in $\mathcal{O}$ are finite

Given A6 and Definition 3, the proof that $< \mathcal{O} \times \mathcal{O}, \geq^d >$ satisfies C5 is trivial and will also be left unstated.

Let me summarise the situation so far. Given S2—we don't need S1 and S3 yet—if $\succsim$ on $\mathcal{O} \cup \mathcal{G}$ satisfies A1 to A6, Theorem 1 implies that there is an effectively unique $U : \mathcal{O} \mapsto \mathbb{R}$ such that:

$$(o_1, o_2) \geq^d (o_3, o_4) \text{ iff } U(o_1) - U(o_2) \geq U(o_3) - U(o_4)$$

Interestingly, we don't yet have enough to derive that $o_1 \succsim o_2$ iff $U(o_1) \geq U(o_2)$, which I take to be a minimal condition on U being a representation of S's preferences over $\mathcal{O}$. To achieve this further result, we need to connect S's preferences over outcomes to the utilities she assigns to the gambles involving those outcomes. We can do this by means of A7 (which will also play an important role later):

Axiom A7 For all $o_1, o_2 \in \mathcal{O}$, $o_1 \succsim o_2$ iff for all $[o_1', P; o_2'] \in \mathcal{G}$, $o_1 \succsim [o_1', P; o_2']$

Essentially: a gamble is never strictly preferred to its best outcome. From this we can also work out that a gamble is never considered to be worse than its worst outcome. (I prove this further below). From this and the earlier axioms, it is then easy to establish that U represents $\succsim$ on $\mathcal{O}$:

Proof From A7, $o_1 \sim o_1'$ iff, for all $[o_1'', P; o_1'''] \in \mathcal{G}$, $o_1 \sim [o_1'', P; o_1''']$. So $o_1 \succsim o_2$ iff $[o_1'', \pi; o_1'''] \succsim [o_2'', \pi'; o_2''']$, which holds iff $(o_1, o_2) \geq^d (o_2, o_1)$. From Theorem 1, $(o_1, o_2) \geq^d (o_2, o_1)$ iff $U(o_1) - U(o_2) \geq U(o_2) - U(o_1)$, which is true just in case $U(o_1) \geq U(o_2)$. So $o_1 \succsim o_2$ iff $U(o_1) \geq U(o_2)$. □

Finally, we will make appeal to one further existential axiom:

Axiom A8 For every $[o_1, P; o_2] \in \mathcal{G}$, there is either (i) an $[o_3, \pi; o_4] \in \mathcal{G}$ such that $[o_1, P; o_2] \sim [o_3, \pi; o_4]$, or (ii) an $o_5 \in \mathcal{O}$ such that $[o_1, P; o_2] \sim o_5$

Note that if the P in $[o_1, P; o_2]$ is in Π, then the first disjunct is trivially satisfied. If it *also* turns out that the second disjunct is satisfied and there's an o_5 such that $[o_1, P; o_2] \sim o_5$, then $U(o_5) = \frac{1}{2} \cdot U(o_1) + \frac{1}{2} \cdot U(o_2)$:

Proof As we've already established, $o_5 \sim [o_5', \pi; o_5'']$, so $[o_5', \pi; o_5''] \sim [o_1, P; o_2]$ (for $P \in \Pi$). This holds just in case $(o_1, o_5) =^d (o_5, o_2)$, from which it quickly follows that $U(o_5) = \frac{1}{2} \cdot U(o_1) + \frac{1}{2} \cdot U(o_2)$. □

Given this, A8 lets us extend U on $\mathcal{O}$ to $\mathcal{O} \cup \mathcal{G}$, by means of the following perfectly reasonable stipulation:

> For all $[o_1, P; o_2] \in \mathcal{G}$: if there is an o_3 such that $[o_1, P; o_2] \sim o_3$, then $U([o_1, P; o_2]) = U(o_3)$; otherwise, if there is an $[o_4, \pi; o_5] \in \mathcal{G}$ such that $[o_1, P; o_2] \sim [o_4, \pi; o_5]$, $U([o_1, P; o_2]) = \frac{1}{2} \cdot U(o_4) + \frac{1}{2} \cdot U(o_5)$

The uniqueness properties of U on $\mathcal{O}$ will carry over to U on $\mathcal{O} \cup \mathcal{G}$.

With the possible exception of A2.2, A8 seems the least plausible of the axioms outlined so far. While it makes intuitive sense for expected utility maximisers with *precise* credences, we should not presume that ordinary agents' credences are always precise. From A1-A7, we get that every *outcome* in $\mathcal{O}$ is assigned a precise utility—which is problem enough, and something we'll have to deal with later on—but A8 adds the additional implication that every *gamble* can be assigned a precise utility. It is reasonable to think, however, that where S's credence towards P is imprecise, her utility towards any (non-trivial) gamble conditional on P should likewise be at least somewhat imprecise. Dealing with imprecision in Cr and U is something we'll come back to in Section 4.2. For now, we'll suppose that A1 to A8 hold.

3.5 Deriving Degrees of Belief

Our goal now is to construct a credence function, $Cr : \mathcal{P} \mapsto [0, 1]$, which combines with U to supply an expected utility representation of $\succsim$ on $\mathcal{O} \cup \mathcal{G}$. As usual, we begin by assuming that Cr will satisfy the properties we intend for it. Suppose that $o_1 \nsim o_2$ and $[o_1, P; o_2] \in \mathcal{G}$. So $U([o_1, P; o_2]) = Cr(P) \cdot U(o_1 \wedge P) + (1 - Cr(P)) \cdot U(o_2 \wedge \neg P)$. Because $U(o_1) = U(o_1 \wedge P) \neq U(o_2) = U(o_2 \wedge \neg P)$, this can be rearranged to provide the following definition of Cr:

Definition 4 (Cr) For all $P \in \mathcal{P}$, if there is a $[o_1, P; o_2] \in \mathcal{G}$ such that $o_1 \nsim o_2$, then:

$$Cr(P) = \frac{U([o_1, P; o_2]) - U(o_2)}{U(o_1) - U(o_2)}$$

A1.2 ensures that there's enough gambles in $\mathcal{G}$ for Definition 4's existential requirements to be satisfied for each $P \in \mathcal{P}$. However, we will also need to make the following rather strong assumption to ensure that Definition 4 is coherent:

Axiom A9 For all $[o_1, P; o_2], [o_3, P; o_4] \in \mathcal{G}$ where $o_1 \nsim o_2$ and $o_3 \nsim o_4$,

$$\frac{U([o_1, P; o_2]) - U(o_2)}{U(o_1) - U(o_2)} = \frac{U([o_3, P; o_4]) - U(o_4)}{U(o_3) - U(o_4)}$$

I am cheating somewhat in stating A9 as I have done here, in terms of U. Since U can be constructed entirely from preferences, A9 is equivalent to *some* condition stated purely in terms of $\succsim$. I do not think, however, that there is much to be gained by rephrasing the axiom in terms of $\succsim$, as doing so would only serve to obscure its content. I'll have more to say about A9 in Section 4.1, specifically about the kinds of representations of S's credences that we might get if we go without it.

For now, what it says can be visualised as follows. Definition 4 tells us that $Cr(P)$ is, say, 0.75, if $o_1 \succ o_2$ and $U([o_1, P; o_2])$ sits exactly three quarters of the way from $U(o_2)$ to $U(o_1)$. In order for Definition 4 to be coherent, therefore, it is important that the value we obtain for $Cr(P)$ does not depend upon the particular choice of gamble $[o_1, P; o_2]$. This is where A9 comes in, requiring that for all pairs of gambles o_1, o_2 such that $o_1 \succ o_2$, if $[o_1, P; o_2] \in \mathcal{G}$, then its utility also sits three quarters of the way from $U(o_2)$ to $U(o_1)$.

This provides us with the resources to prove the following:

Theorem 2 *If S2 and A1-A9 hold, then there is a function U:* $\mathcal{O} \cup \mathcal{G} \mapsto \mathbb{R}$ *and a function* $Cr : \mathcal{P} \mapsto [0, 1]$ *such that for all* $x, y \in \mathcal{O} \cup \mathcal{G}$*, all* $o_1, o_2, o_3, o_4 \in \mathcal{O}$*, all* $[o_1, P; o_2] \in \mathcal{G}$*,*

(i) $x \succsim y$ *iff* $U(x) \geq U(y)$
(ii) $U([o_1, P; o_2]) = Cr(P) \cdot U(o_1) + Cr(\neg P) \cdot U(o_2)$

Furthermore, Cr is unique and U is unique up to positive linear transformation. Additionally, Cr has the property that for all $P \in \mathcal{P}$*,*

(iii) $Cr(P) = 1 - Cr(\neg P)$

Proof That a U satisfying property (i) exists, and that it is unique up to positive linear transformation, has already been established using A1 to A8; the only structural assumption appealed to there was S2. The rest of the proof will proceed in four stages. First, we'll show that Cr is a function from $\mathcal{P}$ into [0, 1]. Then, we will show that property (ii) holds for all $[o_1, P; o_2] \in \mathcal{G}$, and then that (iii) holds for all $P \in \mathcal{P}$. Finally, we will show that Cr is unique.

That Cr maps $\mathcal{P}$ to [0, 1]: That Cr is defined for all $P \in \mathcal{P}$ follows from A1.2 and Definition 4. That $Cr(P)$ is independent of the choice of outcomes and gambles satisfying the antecedent conditions follows immediately from A9. To see that the range of Cr is [0, 1], we'll first show that given A2 and A3, A7 implies that if $o_1 \succsim o_2$, then for all relevant $[o_1', P; o_2']$, $o_1 \succsim [o_1', P; o_2'] \succsim o_2$. That $o_1 \succsim o_2$ implies $o_1 \succsim [o_1', P; o_2']$ is just A7. Next, suppose that $o_1 \succsim o_2$, so either $o_1 \sim o_2$ or $o_1 \succ o_2$. If $o_1 \sim o_2$, then A7 implies that $o_1 \sim [o_1', P; o_2']$ and $o_2 \sim [o_2', \neg P; o_1']$ (recall that $[o_1', P; o_2'] = [o_2', \neg P; o_1']$). If $o_1 \succ o_2$ then $\neg(o_2 \succsim o_1)$, so $\neg(o_2 \succsim [o_2', \neg P; o_1'])$; thus $[o_1', P; o_2'] \succ o_2$. In either case, then, if $o_1 \succsim o_2$ then $o_1 \succsim [o_1', P; o_2'] \succsim o_2$. With that established, we then note that for all $[o_1, P; o_2]$, either $o_1 \succsim o_2$ and $o_1 \succsim [o_1, P; o_2] \succsim o_2$, or $o_2 \succsim o_1$ and $o_2 \succsim [o_1, P; o_2] \succsim o_1$. With the already established properties of U we know $U([o_1, P; o_2])$ sits somewhere weakly between $U(o_1)$ and $U(o_2)$; thus, the difference between $U([o_1, P; o_2])$ and $U(o_2)$ will never be greater than the difference between $U(o_1)$ and $U(o_2)$, and the ratio of those differences will be within [0, 1].

That (ii) holds for all $[o_1, P; o_2] \in \mathcal{G}$: Suppose first that $o_1 \sim o_2$; then, by the reasoning noted above, $U(o_1) = U(o_2) = U([o_1, P; o_2])$. Let $(o_1) = x$, so (ii) holds in this case just in case $x = Cr(P) \cdot x + (1 - Cr(P)) \cdot x$. We have just established that $Cr(P) \in [0, 1]$, so regardless of what value $Cr(P)$ takes the required equality will hold. Suppose next that $o_1 \nsim o_2$. By Definition 4,

$$Cr(P) \;=\; \frac{U([o_1,\ P;\ o_2]) \;-\; U(o_2)}{U(o_1) \;-\; U(o_2)}$$

This holds iff

$$Cr(P) \cdot \big(U(o_1) - U(o_2)\big) \;=\; U([o_1, P; o_2]) - U(o_2)$$

which can be rearranged to

$$U([o_1, P; o_2]) \;=\; Cr(P) \cdot U(o_1) - Cr(P) \cdot U(o_2) + U(o_2)$$

Next we'll demonstrate that $Cr(\neg P) = 1 - Cr(P)$, concluding the proof that (ii) holds.

That (iii) holds for all $P \in \boldsymbol{\mathcal{P}}$: For any pair o_1, o_2 such that $o_1 \nsim o_2$, if $[o_1, P; o_2] \in \boldsymbol{\mathcal{G}}$ then $[o_2, \neg P; o_1] \in \boldsymbol{\mathcal{G}}$. A1.2 ensures that for every $P \in \boldsymbol{\mathcal{P}}$ we'll find the former gamble in $\boldsymbol{\mathcal{G}}$, so we know both are. So:

$$Cr(\neg P) \;=\; \frac{U([o_2,\ \neg P;\ o_1]) \;-\; U(o_1)}{U(o_2) \;-\; U(o_1)}$$

Multiplying the denominator and the numerator by -1 gets us:

$$Cr(\neg P) \;=\; \frac{U(o_1) \;-\; U([o_2,\ \neg P;\ o_1])}{U(o_1) \;-\; U(o_2)}$$

Let $U([o_1, P; o_2]) = x = U([o_2, \neg P; o_1])$. Given the foregoing, $Cr(P) + Cr(\neg P)$ is equal to:

$$\frac{x \;-\; U(o_2)}{U(o_1) \;-\; U(o_2)} + \frac{U(o_1) \;-\; x}{U(o_1) \;-\; U(o_2)} \;=\; \frac{U(o_1) \;-\; U(o_2)}{U(o_1) \;-\; U(o_2)}$$

Thus, $Cr(P) + Cr(\neg P) = 1$, so $Cr(P) = 1 - Cr(\neg P)$.

That Cr is unique: Because ratios of differences will always be preserved across admissible transformations of U, there can be only one Cr such that:

$$Cr(P) \;=\; \frac{U([o_1,\ P;\ o_2]) \;-\; U(o_2)}{U(o_1)\, U(o_2)},\ \text{where} o_1 \nsim o_2 \text{ and } [o_1, P; o_2] \in \boldsymbol{\mathcal{G}}$$

We know that this equality holds iff

$$U([o_1, P; o_2]) \;=\; Cr(P) \cdot U(o_1) + (1 - Cr(P)) \cdot U(o_2)$$

So there is only one Cr that satisfies (ii). □

It's worth pausing to say a few things about the properties of Cr. Perhaps most important, Cr need not be a probability function, nor need it even be monotonic. The main restriction on Cr is that for any $\pi \in \Pi$, $Cr(\pi) = \frac{1}{2}$, and there must be at least two propositions in Π. Also important to note is that none of the propositions in the domain of Cr (or U) need be very *specific*. Because we have placed so few constraints on the internal structures of either $\boldsymbol{\mathcal{O}}$ or $\boldsymbol{\mathcal{P}}$, we need not suppose anything as strong as, say, Jeffrey's assumption that $\boldsymbol{\mathcal{P}}$ (minus a set of 'null propositions') forms a bottomless algebra with ever-increasingly fine-grained contents.

A very simple example will help bring these points out. For simplicity, I'll assume that propositions are coarsely individuated, and that $P \rightharpoonup Q$ iff $P \vdash Q$.

1. $\boldsymbol{\mathcal{P}} = \{\pi, \neg\pi, P, \neg P, P \vee Q, \neg(P \vee Q)\}$
2. $\boldsymbol{\mathcal{O}} = \{o_t, o_m, o_b, o_t \wedge \pi, o_m \wedge \pi, o_b \wedge \pi, o_t \wedge \neg\pi, o_m \wedge \neg\pi, o_b \wedge \neg\pi\}$
3. $o_t \vdash P$ and $o_b \vdash \neg(P \vee Q)$
4. o_t, o_b are logically independent of π
5. o_m is logically independent of every member of $\boldsymbol{\mathcal{P}}$

From S2, we can determine the gambles that will be found in $\mathcal{G}$:

$$[o_t \wedge \pi, \pi; o_t \wedge \neg\pi] = \gamma_1 \quad [o_t \wedge \pi, \pi; o_m \wedge \neg\pi] = \gamma_2 \quad [o_t \wedge \pi, \pi; o_b \wedge \neg\pi] = \gamma_3$$
$$[o_m \wedge \pi, \pi; o_t \wedge \neg\pi] = \gamma_4 \quad [o_m \wedge \pi, \pi; o_m \wedge \neg\pi] = \gamma_5 \quad [o_m \wedge \pi, \pi; o_b \wedge \neg\pi] = \gamma_6$$
$$[o_b \wedge \pi, \pi; o_t \wedge \neg\pi] = \gamma_7 \quad [o_b \wedge \pi, \pi; o_m \wedge \neg\pi] = \gamma_8 \quad [o_b \wedge \pi, \pi; o_b \wedge \neg\pi] = \gamma_9$$
$$[o_t, P; o_b] = \gamma_{10} \qquad [o_t, P \vee Q; o_b] = \gamma_{11}$$

Given a coarse-grained approach to propositions, S3 is satisfied. (If we were to take a very fine-grained approach to propositions, we would need separate entries in $\mathcal{O}$ for o_t, $o_t \wedge P$, $(o_t \wedge P) \wedge P$, and so on; likewise for P, $\neg P$, $\neg\neg P$, etc.).

Assuming that $o_t \succ o_m \succ o_b$ and that $o_i \sim (o_i \wedge \pi) \sim (o_i \wedge \neg\pi)$, A1.1 is also satisfied. π could be thought of as something to the effect of *the next fair coin to be tossed lands heads* (Section 3.3). For every proposition in $\mathcal{P}$ there is a gamble with unequally valued outcomes in $\mathcal{G}$, so A1.2 is satisfied. We now let the following preference relations hold:

$$o_t \sim \gamma_1 \succ \gamma_{2/4/10} \succ o_m \sim \gamma_{3/5/7} \succ \gamma_{6/8/11} \succ o_b \sim \gamma_9$$

It is then easy to check that each of our other axioms are satisfied. Suppose, then, that we normalise our utility function, so that $U(o_t) = 1$, $U(\gamma_{10}) = 0.75$, $U(o_m) = 0.5$, $U(\gamma_{11}) = 0.25$, $U(o_b) = 0$. Given the values for γ_{10} and γ_{11} in particular, Definition 4 implies that $Cr(P) = 0.75$, and $Cr(P \vee Q) = 0.25$. This is despite the fact that, obviously, $P \vdash P \vee Q$. Indeed, $Cr(P)$ and $Cr(P \vee Q)$ could have taken any of a range of other values without falsifying our axioms: the choice to let $\gamma_{10} \sim \gamma_2$ and $\gamma_{11} \sim \gamma_6$ was more or less arbitrary.

The reason for *Cr*'s permissiveness is that *internal* logical relations amongst the members of $\mathcal{P}$ play almost no role in fixing *Cr*'s values. (The only major exception is the requirement that every P be paired with a $\neg P$). Essentially, $\mathcal{P}$ can be thought of as a set of *points*, each of which is related to at least one other point (its negation). The values that *Cr* then attaches to those pairs of points then depends entirely on the placement of gambles involving them, and the axioms leave a great deal of freedom in that respect. In particular, without additional restrictions or assumptions, the axioms we've specified thus far are effectively blind both to the semantic content of $\mathcal{P}$'s members, and to (most of) the logical relations between them. Consequently, there are (almost) no mechanisms by which relations amongst $\mathcal{P}$'s members can be used to determine the relative placement of gambles within the preference ordering.

Theorem 2 is thus compatible with a wide range of credence functions. Indeed, *Cr* is capable of assigning values of greater than 0 to impossible propositions, and less than 1 to necessary propositions. In the above model add '$P \vee \neg P$' and '$P \wedge \neg P$' to $\mathcal{P}$, and to $\mathcal{O}$ add $o_n = P \wedge \neg P$, with $o_n \sim o_b$. As every consistent outcome in $\mathcal{O}$ implies $P \vee \neg P$ but only o_n implies $P \wedge \neg P$, we'll have nine new gambles in $\mathcal{G}$. Three in particular are salient:

$$[o_t, P \vee \neg P; o_n] = \gamma_{12} \quad [o_m, P \vee \neg P; o_n] = \gamma_{13} \quad [o_b, P \vee \neg P; o_n] = \gamma_{14}$$

From A7, we know that $\gamma_{14} \sim o_b$, but we have a lot more freedom with respect to the placement of γ_{12} and γ_{13}. If $\gamma_{12} \sim o_t$, then $Cr(P \vee \neg P) = 1$. If $\gamma_{12} \sim o_m$,

then $Cr(P \vee \neg P) = 0.5$ (and $P \vee \neg P \in \Pi$). If $\gamma_{12} \sim o_b$, then $Cr(P \vee \neg P) = 0$. Of course, on a coarse-grained approach to thought content, then no one *plausibly* attaches zero credence to $P \vee \neg P$ (or has preferences like $\gamma_{12} \sim o_b$ and $\gamma_{12} \sim o_m$), but we don't need our axioms to preclude that possibility either.

I take the lack of imposed structure on Cr to be a feature, not a bug. Plausibly, ordinary agents don't have probabilistically coherent degrees of belief, so any representation of our credences which *presupposes* such coherence is flawed. (And nothing I've said implies that *Cr can't* be coherent). In Section 3.6, I will suggest a further axiom which ensures that $Cr(P) = Cr(Q)$ just in case $P \rightleftharpoons Q$. With that condition in place, we can reasonably expect that any blatant impossibilities are assigned a credence of 0, and any obvious logical necessities a credence of 1. With additional axioms, it's also possible to ensure that Cr satisfies particular structural properties, such as a weakened form of monotonicity: if $P \rightharpoonup Q$, then $Cr(Q) \geq Cr(P)$ (see [15, Section 4.4]).

3.6 Indifference Under $\rightleftharpoons$ Equivalence

We have not said anything yet to guarantee that $U(o_1) = U(o_1 \wedge P)$, and this is a problem. Take a pair of propositions o_1 and P such that $o_1 \rightharpoonup P$. Supposing that all of the axioms so far stated are satisfied, Theorem 2 tells us that we can derive some Cr and U such that for any $[o_1, P; o_2] \in \mathcal{G}$, $U([o_1, P; o_2]) = Cr(P) \cdot U(o_1) + Cr(\neg P) \cdot U(o_2)$. This the right result only if $U(o_1) = U(o_1 \wedge P)$ and $U(o_2) = U(o_2 \wedge \neg P)$. Where this is not the case, we should really be considering our utilities for the outcomes *under the conditions that obtain if they are won*. This, I take it, should be taken as both a normative requirement and a plausible descriptive condition: ordinary agents do not ignore background conditions when evaluating an outcome. However, it is entirely consistent with A1 to A9 that $o_1 \nsim (o_1 \wedge P)$, and so $U(o_1) \neq U(o_1 \wedge P)$. In this case, Theorem 2 would have us represent S's utilities over gambles in a way that only seems appropriate if $o \rightharpoonup P$ implies $o \sim (o \wedge P)$, but it achieves this *via* a set of axioms which are consistent with $o \rightharpoonup P$ and $o \nsim (o \wedge P)$. Something has gone wrong.

A condition that we could add to avoid this kind of situation would be:

Axiom 10a For all o $\in \mathcal{O}$ and $P \in \mathcal{P}$, if $o \rightharpoonup P$, then $o \sim (o \wedge P)$

In the paragraphs that follow, I want to argue that we ought to posit something stronger than this. The key issue concerns what kind of *justification* we could have for A10a, or any other axiom we might add to play the same role.

Much here depends on how we interpret '$\rightharpoonup$'. That relation hasn't played any interesting role in the foregoing proofs. In particular, we've not yet made use of S1, so for the purely formal purposes of proving Theorem 2, $\rightharpoonup$ might as well be any arbitrary relation between outcomes and propositions. Soon, we'll use S1 in establishing that if $o \rightharpoonup P$, then $U(o) = U(o \wedge P)$. But there are plenty of relations which satisfy the conditions of S1; the most obvious are various kinds of *necessitation* or *implication* relations. What we need, therefore, is an interpretation of '$\rightharpoonup$' which

both satisfies S1 and would justify something like A10a. I've already suggested three possibilities, which I'll return to shortly below. But first, it will be fruitful to consider some interpretations which *won't* give us what we want.

Pick your favourite sense of 'necessitates': logical, metaphysical, or *a priori*; then suppose that '$o \rightharpoonup P$' means *o necessitates P*. Under this interpretation, $\mathcal{G}$ will consist of all and only the two-outcome gambles that can be constructed from $\mathcal{O}$ and $\mathcal{P}$ such that the outcomes necessitate the conditions under which they are won. Axiom A10a then asserts that for pairs of outcome-propositions o and $o \wedge P$ such that o necessitates P (so necessarily, o and $o \wedge P$ are necessarily equivalent), $o \sim (o \wedge P)$. Now, this isn't quite as strong as saying that S is indifferent between *all* pairs of necessarily equivalent propositions—but it's hard to see how A10a could be justified in any way without also justifying the stronger claim. After all, it imposes a very strict kind of inferential infallibility upon the agent—an ability to always recognise whenever o necessitates P, for any pair of o in $\mathcal{O}$ and P in $\mathcal{P}$. There seems to be no important difference between this kind of infallibility and the more general ability to determine the relevant necessitation relationships between *any* arbitrary pair of propositions that might be considered. So it seems that if '$o \rightharpoonup P$' is taken to mean that *o necessitates P* in any of those strong senses, then A10a looks pretty implausible for ordinary agents. The representations we arrive at with Theorem 2 are then only plausible for certain kinds of idealised agent who always assign the same utilities to equivalent propositions. We are left without a representation for the average person on the street, who lacks such intellectual brilliance and who may fail to *recognise* that o necessitates P.

Furthermore, the remarked upon flexibility of Cr becomes rather odd under this kind of interpretation of $\rightharpoonup$. If our subject is always able to recognise implication relations, then we might expect her credences to satisfy *at least* a basic monotonicity condition: if P necessitates Q, then $Cr(P) \leq Cr(Q)$. However, we have seen that (without further axioms) Cr need not be monotonic. Indeed, the joint satisfaction of each of the axioms outlined so far (including Axiom A10a) is compatible with $Cr(P) \neq Cr(Q)$ for equivalent P and Q. It seems implausible to demand of an agent extraordinary intellectual capabilities with respect to one domain (U on $\mathcal{O}$), whilst at the same time representing that agent as highly irrational with respect to another domain (Cr on $\mathcal{P}$). Inasmuch as we need to presuppose that S has some special kind of inferential capacity to make Theorem 2's U a plausible representation of S's utilities, it had better not be the case that the very same theorem supplies us with a credence function which would only make sense when taken to represent the inferentially incompetent!

What the foregoing suggests, then, is that we require something stronger than A10a. We need an axiom which will impose a greater degree of consistency in S's credences and utilities in the event that $P \rightleftharpoons Q$. To that end, the following axiom looks plausible:

Axiom A10 For all $o_1, o_2 \in \mathcal{O}$, if $o_1 \rightleftharpoons o_2$, then $o_1 \sim o_2$; and for all $P, Q \in \mathcal{P}$, if $P \rightleftharpoons Q$, then if $[o_1, P; o_2], [o_1', Q; o_2'] \in \mathcal{G}$, then $[o_1, P; o_2] \sim [o_1', Q; o_2']$

In light of the other axioms, A10 implies A10a, and ensures that if P and Q are in the domain of Cr and/or U and $P \rightleftharpoons Q$, then $Cr(P) = Cr(Q)$ and $U(P) = U(Q)$:[14]

Proof Given the established properties of U, if $o_1, o_2 \in \mathcal{O}$ and $o_1 \sim o_2$, then $U(o_1) = U(o_2)$; so, if $o_1 \rightleftharpoons o_2$ and A10 holds, then $U(o_1) = U(o_2)$. Likewise, given the established properties of Cr, if $P, Q \in \mathcal{P}$ and $[o_1, P; o_2], [o_1', Q; o_2'] \in \mathcal{G}$, then A10 states that if $[o_1, P; o_2] \sim [o_1', Q; o_2']$, then $Cr(P) = Cr(Q)$. So we just need to show that the right pair of gambles exists in $\mathcal{G}$. For this we use S1.2 and S1.3. Given that $P \rightleftharpoons Q$ implies that $o \rightharpoonup P$ iff $o \rightharpoonup Q$, and that $P \rightleftharpoons Q$ iff $\neg P \rightleftharpoons \neg Q$, if some $[o_1, P; o_2]$ is in $\mathcal{G}$ then $[o_1, Q; o_2]$ is certain to be in $\mathcal{G}$ as well. Thus, if $P \rightleftharpoons Q$ and $P, Q \in \mathcal{P}$, $Cr(P) = Cr(Q)$. □

On the other hand, if $\neg(P \rightleftharpoons Q)$, then $Cr(P)$ need not equal $Cr(Q)$, and $U(P)$ need not equal $U(Q)$. P and Q *may* still be assigned the same values by Cr and U, but it won't fall out of the theorem that they must be. Furthermore, given S3 and S1.1's implication that $o_1 \rightleftharpoons (o_1 \wedge P)$ whenever $o_1 \rightharpoonup P$, A10 also guarantees that $U(o_1) = U(o_1 \wedge P)$.

Now, A10 seems like an incredibly plausible assumption—both normatively and descriptively—under either Interpretation 1 (believed implication), Interpretation 2 (recognised implication), or Interpretation 3 (obvious implication). If we take the most flexible interpretation, Interpretation 1, A10 says that if S believes that P and Q are just the same state of affairs, then (i) she will be indifferent between P and Q, and (ii) she will also be indifferent between any two gambles of the form $[o_1, P; o_2]$ and $[o_1', Q; o_2']$. After all, each has a $Cr(P) = Cr(Q)$ likelihood of resulting in an outcome equal in value to o_1 and a $1 - Cr(P)$ likelihood of resulting in an outcome equal in value to o_2. That is, she does not distinguish between purportedly equivalent propositions P and Q when forming her preferences over $\mathcal{O}$ and $\mathcal{G}$. Where P and Q are *recognised* as being equivalent, or even *obviously equivalent*, we should expect A10 to hold all the more so. Even if ordinary agents don't live up to this very weak standard of rationality, it can hardly be doubted that they *approximate* the condition quite closely—and any agent who does not even come close to satisfying A10 (on any of the three suggested interpretations) is likely too irrational to have coherently measured credences and utilities in any case.

We now have enough to state our main representation result:

[14]Note the role of S1.2 and S1.3 in this proof: they are used to establish that if $P \rightleftharpoons Q$, then if $[o_1, P; o_2]$ is in $\mathcal{G}$, $[o_1, Q; o_2]$ will be in $\mathcal{G}$ too. Given A9, we could get away with dropping both conditions if we made the relatively weak partially structural assumption that when $P \rightleftharpoons Q$, there is *some* pair of outcomes o_1, o_2 such that there are $[o_1, P; o_2]$, $[o_1', Q; o_2'] \in \mathcal{G}$, and $[o_1, P; o_2] \sim [o_1', Q; o_2']$. Alternatively, we could tweak the second part of A10 to say that if $P \rightleftharpoons Q$, then there will be a pair of gambles $[o_1, P; o_2]$, $[o_3, Q; o_4]$ in $\mathcal{G}$ such that:

$$\frac{U([o_1,\ P;\ o_2]) - U(o_2)}{U(o_1) - U(o_2)} = \frac{U([o_3,\ Q;\ o_4]) - U(o_4)}{U(o_3) - U(o_4)}$$

However, this latter option would result in an axiom somewhat less intuitive than A10 as stated. Finally, if we wanted to get rid of S1.2, S1.3 *and* A9 while preserving the result that $P \rightleftharpoons Q$ implies $Cr^*(P) = Cr^*(Q)$ (see Section 4.1), we'd need to posit that whenever $P \rightleftharpoons Q$ and $[o_1, P; o_2] \in \mathcal{G}$, there's a $[o_1', Q; o_2']$ in $\mathcal{G}$.

Theorem 3 *If S1-S3 and A1-A10 hold, then there is a function U: $\mathcal{O} \cup \mathcal{G} \mapsto \mathbb{R}$ and a function Cr : $\mathcal{P} \mapsto [0, 1]$ such that for all $x, y \in \mathcal{O} \cup \mathcal{G}$, all $o_1, o_2, o_3, o_4 \in \mathcal{O}$, all $[o_1, P; o_2] \in \mathcal{G}$, and all $P \in \mathcal{P}$,*

(i) $x \succsim y$ *iff* $U(x) \geq U(y)$
(ii) $U([o_1, P; o_2]) = Cr(P) \cdot U(o_1 \wedge P) + Cr(\neg P) \cdot U(o_2 \wedge \neg P)$

Furthermore, Cr is unique and U is unique up to positive linear transformation. Additionally, Cr and U have the properties that for all $P, Q \in \mathcal{P}$ and $o_1, o_2 \in \mathcal{O}$,

(iii) $Cr(P) = 1 - Cr(\neg P)$
(iv) *If* $P \rightleftharpoons Q$, *then* $Cr(P) = Cr(Q)$, *and if* $o_1 \rightleftharpoons o_2$, *then* $U(o_1) = U(o_2)$

Proof The proof has already been given that if $P \rightleftharpoons Q$, then $Cr(P) = Cr(Q)$ and $U(P) = U(Q)$, for relevant $P, Q \in \mathcal{P} \cup \mathcal{O}$. That $o_1 \wedge P \in \mathcal{O}$ whenever $[o_1, P; o_2] \in \mathcal{G}$ is asserted directly by S3, so we know $U(o_1 \wedge P)$ and $U(o_2 \wedge \neg P)$ are both defined. As we already know that $U([o_1, P; o_2]) = Cr(P) \cdot U(o_1) + Cr(\neg P) \cdot U(o_2)$ when S2-S3 and A1-A9 hold, and since A10 and S1 jointly imply that $U(o_1) = U(o_1 \wedge P)$ and $U(o_2) = U(o_2 \wedge \neg P)$, property (ii) follows immediately. □

The way we interpret '$\rightharpoonup$' thus makes a significant difference to how we interpret Theorem 3 as a whole. What is the *best* way to interpret '$\rightharpoonup$'? Well, that depends on your aims. If you're interested in a certain kind of idealised agent, you might like to read it as a kind of epistemic necessitation relation. My own aim is to see how much information about an ordinary subject's doxastic states can be extracted from her preferences under certain reasonably minimal assumptions about her. For that reason, my first preference is for Interpretation 1.

To be sure, there is a *prima facie* tension with adopting that interpretation given my aims: any specification of $\rightharpoonup$ under Interpretation 1 requires epistemic access to when *S believes* that *P* implies *Q*. But if I needed that kind of access to *S*'s doxastic states before making use of the theorem, I'd probably be helping myself to too much. But there are options here for the fan of Interpretation 1. In particular, we can let the specification of when $\rightharpoonup$ holds be one more variable to be fixed (or at least significantly restricted) by the information we have of *S*'s preferences.

For instance, if we grant that $\rightharpoonup$ generally satisfies S1.1 and that ordinary agents generally satisfy A10, then for any *plausible* specification of $\rightharpoonup$ (and therefore $\mathcal{G}$), we should expect to find the following:

> If $P \rightharpoonup Q$, then (i) $P \sim (P \wedge Q)$; and (ii) for any pair of gambles $[o_1, P; o_2]$, $[o_1', P \wedge Q; o_2']$ in $\mathcal{G}$, $[o_1, P; o_2] \sim [o_1', P \wedge Q; o_2']$

We can take this expectation to generate a basic criterion of adequacy for any potential specification of when $\rightharpoonup$ holds between two propositions. Roughly: *S* probably doesn't believe that *P* implies *Q* if she doesn't have the kinds of preference patterns we'd expect of someone with those beliefs. If nothing else, this gives us a plausible way of ruling out certain hypotheses about whether *S* believes *P* implies *Q*.

I suspect that this won't be enough to pin down $\rightharpoonup$ exactly: there may be many possible ways of specifying $\rightharpoonup$ consistent with the criterion's satisfaction. To help

deal with some of the left-over indeterminacy, something like a principle of charity would likely be useful. If we have a default assumption that agents' beliefs about whether P implies Q are generally accurate, then we have a secondary way of filtering out hypotheses about when $\rightharpoonup$ does and doesn't hold. It's also worth emphasising that we don't need $\succsim$ to determine Cr and U down to complete uniqueness for the kinds of purposes I have in mind (Section 2), and it's probably a mistake to expect that it should. It is enough to show that information about S's preferences imposes very significant and interesting constraints on what credences and utilities she might have.

Interpretation 2 presents a similar kind of 'access' worries to Interpretation 1—recognition is a species of belief, and we certainly can't expect ordinary agents to recognise arbitrary implication relationships whenever they hold. However, because recognition is factive, we can expect to get a little more purchase on when S recognises that P implies Q from her preferences. In particular, we can modify the above criterion of adequacy to get the following:

> If $P \rightharpoonup Q$, then (i) P *implies* Q, (ii) $P \sim (P \wedge Q)$; and (iii) for any pair of gambles $[o_1, P; o_2]$, $[o_1', P \wedge Q; o_2']$ in $\boldsymbol{\mathcal{G}}$, $[o_1, P; o_2] \sim [o_1', P \wedge Q; o_2']$

Indeed, if P actually implies Q, *and* S's preferences are *as if* she recognises this fact, then that is certainly strong evidence that she does recognise that P implies Q.

Finally, there is also the *obvious implication* interpretation. For this I have in mind a kind of *objective* obviousness. I take it that there are clear cases in which P obviously implies Q. For instance, *there are dogs* obviously implies that *there are things*; and *there are dogs and cats* obviously implies *there are cats*. And there are clear cases where P non-obviously implies Q. For instance, *there are dogs* implies that *there are infinitely many primes*, but this is by no means obvious. But, between these, there are also cases where an implication may be obvious to some, but not so obvious to others. I don't see any benefit to adopting this subjective notion of obviousness in the interpretation of '$\rightharpoonup$', as it presents the same kinds of 'access' difficulties that Interpretation 1 and Interpretation 2 have without any clear additional benefits. On the other hand, suppose there are some inferences which just *are* obvious to everyone in your community, which everyone recognises as obvious, and which *should* be obvious to anyone worthy of being called an agent, at least in *normal conditions*. We would then have a way of interpreting '$\rightharpoonup$' as a restriction on the ordinary implication relation which we can plausibly assume S reliably reasons in accord with, without ever needing to peek inside S's head.

There are two main difficulties with adopting Interpretation 3. The first (and most obvious!) is in spelling out the conditions for when P objectively obviously implies Q. I do not know how this would be done. One might appeal to a notion of minimal rationality as a constitutive norm of agency: it's not at all implausible that part of *what it is* for S to be an agent at all is for S to be *minimally rational*, to respond appropriately to the evidence around her, and to make rational choices in light of that evidence. To say that S is an agent is to presuppose that S at least comes close to satisfying some basic criterion of rationality. If so, it would be natural to presume that S draws out any obvious implications from the propositions she considers, and recognises obvious logical equivalences, *ceterus paribus*.

The second difficulty is more worrying to me. It's plausible that for any P, there are very few things that P objectively obviously implies. If so, Interpretation 3 generates significant tension with the main existential assumption we've made about gambles, A1. The more tightly restricted $\rightharpoonup$ is, the less plausible this axiom will be—and if there are too few gambles in $\mathcal{G}$, we won't get very much of a fix on Cr and U. The other two interpretations are less restrictive, which is an important benefit that they have over Interpretation 3.

4 Weakening the Axioms

In this section, I want to discuss three independent ways of modifying the axioms outlined in Section 3. In particular, I want to see what can be done without the strong consistency axiom A9 (Section 4.1), and how we can modify the system to accommodate imprecise credences and utilities by getting rid of A8 and A2.2 (Section 4.2).

4.1 Accommodating Inconsistency

A9 is a very strong condition. I'm inclined to think that it is the least plausible of the axioms outlined in this paper. It essentially requires of our subject an extraordinary degree of *precision* and *consistency* with respect to how she ranks different gambles conditional on the same proposition. It is the kind of axiom which would only make sense of subjects who are unflinchingly consistent *qua* expected utility maximisers with precise credences, always calculating their value precisely according to the expected utility formula. This is too much to ask of ordinary agents, who we should at best only expect to approximate such norms of consistency.

It's also worth noting that A9 also ends up implying analogues of controversial independence axioms from other expected utility theorems. For instance, it implies something much like von Neumann & Morgenstern's [42] 'Independence' axiom: if $o_1 \succsim o_2$, then for any pair $[o_1', P; o_3], [o_2', P; o_3'] \in \mathcal{G}$, $[o_1', P; o_3] \succsim [o_2', P; o_3']$. Likewise, it implies restricted analogues of Savage's axioms P2 and P4: (i) for all relevant gambles in $\mathcal{G}$, $[o_1, P; o_2] \succ [o_3, P; o_2]$ if and only if $[o_1, P; o_4] \succ [o_3, P; o_4]$; and (ii) if $o_1 \succ o_2$ and $[o_1, P; o_2] \succsim [o_1', Q; o_2']$, then for all relevant gambles in $\mathcal{G}$, if $o_3 \succ o_4$, then $[o_3, P; o_4] \succsim [o_3', Q; o_4']$. It would certainly be best to do without these implications, to whatever extent possible.

As it turns out, we can to some extent do without A9 by making some tweaks to our definition—and interpretation—of Cr:

Definition 5 (Cr^*) For all $P \in \mathcal{P}$, $Cr^*(P) = [\lambda_1, \lambda_2]$ iff $[\lambda_1, \lambda_2]$ is the smallest interval such that for any $[o_1, P; o_2] \in \mathcal{G}$ where $o_1 \nsim o_2$,

$$\frac{U([o_1,\ P;\ o_2])\ -\ U(o_2)}{U(o_1)\ -\ U(o_2)} \in [\lambda_1, \lambda_2]$$

For more or less the same reason that Cr is unique, Cr^* will also be unique. In what follows, I'll first say a few words about Definition 5 and the results we get if we drop A9, after which I'll discuss how I think Cr^* ought to be interpreted.

Here is the intuitive idea behind Cr^*. Definition 4 essentially says that $Cr(P) = \frac{1}{n}$ just in case the agent treats *all* gambles conditional on P precisely as though she assigns a credence of $\frac{1}{n}$ to P, in the sense that the value of U for all gambles $[o_1, P; o_2]$ with $o_1 \succ o_2$ sit $\frac{1}{n}$ of the way between o_1 and o_2. Definition 5, on the other hand, allows for some variability in the agent's preferences with respect to gambles conditional on P, and Cr^* represents that variation by means of an interval. For example, suppose (i) that $o_1 \succ o_2 \succ o_3 \succ o_4$, (ii) the agent's utility for $[o_1, P; o_2]$ sits $\frac{1}{2}$ way between her utilities for o_1 and o_2, (iii) her utility for $[o_3, P; o_4]$ is $\frac{1}{4}$ of the way between her utilities for o_3 and o_4, and finally (iv) $[o_1, P; o_2]$ and $[o_3, P; o_4]$ are the only gambles conditional on P. Then, $Cr^*(P)$ would be $[\frac{1}{4}, \frac{1}{2}]$. If there were yet more gambles to consider—say, $[o_1, P; o_4]$—and its value sat $\frac{1}{3}$ of the way between the values of its outcomes, then $Cr^*(P)$ would remain unchanged; however, if it was $\frac{1}{5}$ of the way, then $Cr^*(P)$ would equal $[\frac{1}{5}, \frac{1}{2}]$. Notice, though, that if A9 *is* satisfied, then $Cr(P) = n$ just in case $Cr^*(P) = [n, n]$. Thus, Cr^* can be seen as a generalisation of Cr, the latter reducing to the former in the special case that A9 holds.

Theorem 4 *If S1-S3 and A1-A8, A10 hold, then there is a function* $U : \mathcal{O} \cup \mathcal{G} \mapsto \mathbb{R}$ *and a function* $Cr^* : \mathcal{P} \mapsto \{[\lambda_1, \lambda_2] : [\lambda_1, \lambda_2] \subseteq [0, 1]\}$ *such that for all* $x, y \in \mathcal{O} \cup \mathcal{G}$*, all* $o_1, o_2, o_3, o_4 \in \mathcal{O}$ *and all* $P \in \mathcal{P}$*,*

(i) $x \succsim y$ *iff* $U(x) \geq U(y)$
(ii)* $U([o_1, P; o_2]) = \lambda \cdot U(o_1) + (1 - \lambda) \cdot U(o_2)$*, for some* $\lambda \in Cr^*(P)$

Furthermore, there is only one Cr^* *satisfying Definition 5, and U is unique up to positive linear transformation. Additionally,* Cr^* *and U have the properties that for all* $P, Q \in \mathcal{P}$ *and* $o_1, o_2 \in \mathcal{O}$*,*

(iii)* $Cr^*(P) = [\lambda_1, \lambda_2]$ *iff* $Cr^*(\neg P) = [1 - \lambda_2, 1 - \lambda_1]$
(iv)* *If* $P \rightleftharpoons Q$*, then* $Cr^*(P) = Cr^*(Q)$*, and if* $o_1 \rightleftharpoons o_2$*, then* $U(o_1) = U(o_2)$

The proof of Theorem 4 essentially reiterates the proof of Theorem 3, and so has been suppressed. The only important difference between the two is that instead of using A9 to justify deriving the value and properties of $Cr(P)$ from an arbitrarily chosen gamble $[o_1, P; o_2]$ where $o_1 \nsim o_2$, we instead perform the same steps on the lowest and highest ranked gambles conditional on P (with differently valued outcomes) to derive the lower and upper bounds of $Cr^*(P)$ as defined by Definition 5. Basically the same steps that were used to establish (ii), (iii) and (iv) in the proof of Theorem 2 can then be used to show that properties (ii*), (iii*) and (iv*) hold.

I *don't* think it would be right to interpret Cr^* as a representation of S's credences *per se*. Rather, the most plausible way to understand Cr^* is to take it as providing us with a *limit* on any adequate measure of S's credences towards P, *given* the information we have about her preferences and assuming that she at least approximately evaluates the utility of gambles according to their expected utility. In other words, I

would suggest that $Cr^*(P) = [\lambda_1, \lambda_2]$ tells us that the agent's preferences constrain what her credence in P are likely to be at least down to $[\lambda_1, \lambda_2]$, on the presupposition that she generally approximates the norm of expected utility maximisation.

This way of reading of Cr^* is compatible with a range of possibilities. For instance, $Cr^*(P) = [\lambda_1, \lambda_2]$ would be consistent with the agent having a precise credence for P anywhere within $[\lambda_1, \lambda_2]$—in which case she is presumably somewhat inconsistent with respect to how she evaluates the utilities of gambles conditional on P. It is also compatible with the agent having imprecise credences accurately measured by some interval within $[\lambda_1, \lambda_2]$, including but not necessarily $[\lambda_1, \lambda_2]$ itself. In either case, though, $Cr^*(P) = [\lambda_1, \lambda_2]$ should only be taken to mean that *whatever* the true measure of the agent's credences regarding P may be, it (most likely) sits *somewhere* within $[\lambda_1, \lambda_2]$. Further information would need to be considered to determine how *exactly* the agent's credences in P should be represented.

4.2 Accommodating Imprecision

Ordinary agents probably don't have infinitely precise credences and utilities, so it's probably not a good idea to represent their credences and utilities using real-valued Cr and U. Real-valued functions imply *comparative completeness*: for all o_1 and o_2 in $\mathcal{O}$, either $U(o_1) \geq U(o_2)$ or $U(o_2) \geq U(o_1)$; and for all P and Q in $\mathcal{P}$, either $Cr(P) \geq Cr(Q)$ or $Cr(Q) \geq Cr(P)$. But it seems reasonable to expect that there will be some incompleteness in our preferences, and likewise some incompleteness in our overall confidence ranking. Let '$P \succsim^b Q$' mean that S takes P to be *at least as probable as* Q. Then we should allow both $\succsim$ and $\succsim^b$ to be incomplete, and Cr and U should reflect this.

Since the Ramseyean strategy is to derive credences from utilities, and utilities from preferences, I'll start with a look at accommodating imprecise credences under the continued assumption that utilities are precise. In Section 3.4, I said that it was plausible that where S's credence in P is imprecise, her utility towards any interesting gamble conditional on P should likewise be at least somewhat imprecise.[15] If this is correct, then A8 (in the context of the other axioms) stands out as particularly implausible for any agent with imprecise credences. Where every outcome in $\mathcal{O}$ is assigned a precise utility, if the utility for $[o_1, P; o_2]$ is imprecise because S's credence for P is imprecise, then there will be no $o_3 \in \mathcal{O}$ such that $[o_1, P; o_2] \sim o_3$; nor will there be a gamble conditional on a $\frac{1}{2}$-probability proposition equipreferable

[15] How agents make decisions with imprecise credences is a matter of much contemporary discussion, so I cannot say anything very definite here. For an overview of the main approaches to decision-making with imprecise credences, see [39]. For a very natural model according to which imprecise credences will generate imprecise utilities for gambles, see [35]. Most descriptively-motivated models of decision-making with imprecise credences aim at representing the apparently risk-averse attitudes that ordinary subjects take towards gambles conditional on propositions with 'ambiguous' probabilities. As such, it is unclear how well they fit with the assumption that our subject is risk-neutral (Section 3.1). If it turns out that otherwise ordinary, risk-neutral agents with imprecise credences follow a rule quite unlike expected utility maximisation, like Γ-maximin (see [1, 38]), then Theorem 3 will likely have to be revised at a very fundamental level (e.g., the motivations for Definitions 1, 2 and 4 will be undermined).

to $[o_1, P; o_2]$.[16] Let $\mathcal{G}^{\Pi}$ be that subset of gambles in $\mathcal{G}$ conditional on $\frac{1}{2}$-probability propositions. Then, the most we should expect is:

Axiom A8* For each $[o_1, P; o_2] \in \mathcal{G}$, there are some $x, y \in \mathcal{O}\mathcal{G}^{\Pi}$ such that $x \succsim [o_1, P; o_2] \succsim y$

Suppose we replace A8 with A8*, and allow for some incompleteness in $\succsim$ specifically with respect to gambles *outside* of $\mathcal{G}^{\Pi}$. (That is, if there is a $z \in \mathcal{O} \cup \mathcal{G}^{\Pi}$ valued between x and y for $x \succsim [o_1, P; o_2] \succsim y$, then it's possible for $[o_1, P; o_2]$ to be incomparable with z). Without the stronger axiom A8, it won't be possible to assign every gamble a precise utility. By itself, A7 ensures that $o_1 \succsim [o_1, P; o_2] \succsim o_2$, but that's not much to go on when it comes to fixing $Cr(P)$. In many cases, however, we'll be able to get more fine-grained than this. For any o_3 and o_4 such that $o_1 \succsim o_3 \succsim [o_1, P; o_2] \succsim o_4 \succsim o_2$, we'll know that the utility of $[o_1, P; o_2]$ (whether precise or not) must be somewhere weakly between $U(o_3)$ and $U(o_4)$. That's a start.

Make the simplifying assumption that for every gamble $[o_1, P; o_2]$, there is an l and an r in $\mathcal{O} \cup \mathcal{G}^{\Pi}$ such that (i) $l \succsim [o_1, P; o_2]$ and there is no $x \in \mathcal{O} \cup \mathcal{G}^{\Pi}$ such that $l \succ x \succsim [o_1, P; o_2]$, and (ii) $[o_1, P; o_2] \succsim r$ and there is no $y \in \mathcal{O} \cup \mathcal{G}^{\Pi}$ such that $[o_1, P; o_2] \succsim y \succ r$. In other words, l and r are the (or amongst the) 'closest' two outcomes on the left and right sides of $[o_1, P; o_2]$ respectively. (A little extra footwork is involved if there are no closest outcomes, but the points that follow won't be changed much if so). We know from A7 that if $o_1 \succsim o_2$, then $o_1 \succsim l \succsim r \succsim o_2$; and if $o_2 \succsim o_1$, then $o_2 \succsim l \succsim r \succsim o_1$. We can then let Cr^{+} be defined as follows:

Definition 6 ($\boldsymbol{Cr^{+}}$) For all $P \in \mathcal{P}$, if there is a gamble $[o_1, P; o_2] \in \mathcal{G}$ such that $o_1 \nsim o_2$, then if $l \sim r$,

$$Cr^{+}(P) = \frac{U(l) \; - \; U(o_2)}{U(o_1) \; - \; U(o_2)}$$

Otherwise, if $l \succ r$, then $Cr^{+}(P) = (\lambda_1, \lambda_2)$, where:

$$\lambda_1 = \frac{U(l) \; - \; U(o_2)}{U(o_1) \; - \; U(o_2)}, \quad \lambda_2 = \frac{U(r) \; - \; U(o_2)}{U(o_1) \; - \; U(o_2)}$$

If $l \sim r$, then $Cr^{+}(P) = Cr(P)$. When $l \succ r$, the gamble $[o_1, P; o_2]$ will have no precise utility, but its location within the preference ranking will be represented by the open interval $(U(l), U(r))$, for which $U(l) = \lambda_1 \cdot U(o_1) + (1 - \lambda_1) \cdot U(o_2)$, and $U(r) = \lambda_2 \cdot U(o_1) + (1 - \lambda_2) \cdot U(o_2)$. To ensure $Cr^{+}(P)$ is well-defined, we'd need to make corresponding changes to A9—i.e., we'll want to ensure that the value of $Cr^{+}(P)$ is independent of whatever gamble conditional on P is chosen, so long as it has outcomes of unequal value. Alternatively, we can simply remove A9 as we did in Section 4.1, and construct a still more complicated function on $\mathcal{P}$.

[16] With the exception of A9—which I'll return to in a moment—none of the other axioms appear especially problematic if S has imprecise credences. A1.1 does presuppose that there are at least two propositions towards which S assigns a credence of exactly $\frac{1}{2}$, but this seems a relatively minor idealisation.

It would be too hasty to read Cr^+ as a direct representation of S's credences; without further conditions, it should be given an interpretation similar to the one suggested for Cr^*. The reason for this is that nothing in our axioms ensures that U is continuous: between any two outcomes o_1 and o_2 such that $o_1 \succ o_2$, there may be no third outcome o_3 such that $o_1 \succ o_3 \succ o_2$. As such, there will always be two (not mutually exclusive) explanations for whenever we find that $l \succ [o_1, P; o_2] \succ r$: either S's credences towards P are imprecise, or there simply weren't enough outcomes between l and r to pin down her actual utility for $[o_1, P; o_2]$. Again: Cr^+ expresses a *limit* on what S's credences might be, given some background assumptions. Even if it doesn't nail down her credences exactly, there's nothing wrong with a little indeterminacy—so long as we're not trying to construct S's credences entirely out of her preferences. And, crucially, we've done nothing to *presuppose* that S's credences must in general be precise.

Finally, we need to accommodate incomplete preference orderings. About this I will be brief, as the general strategy I'll mention is fairly well known (see, e.g., [3, 19, 33, 44]). First, for any $\succsim$ on $\mathcal{O} \cup \mathcal{P}$, call $\succsim^+$ a *completion* of $\succsim$ just in case (i) $\succsim^+$ is complete and (ii) if $x \succsim y$, then $x \succsim^+ y$. Second, say that $\succsim$ is *coherent* just in case it satisfies A1 to A7. And third, if $\succsim$ is incomplete, say that it is *coherently completable* just in case there is at least one coherent completion of $\succsim$. In Section 3.4 we proved that every coherent $\succsim$ could be represented by a real-valued function U. The natural thing to do, then, when representing an incomplete but coherently completable preference ordering, is to have U map each $o \in \mathcal{O}$ to the set of values it might take under each possible coherent completion of $\succsim$. [17]

Requiring only that $\succsim$ should be coherently completable significantly weakens the requirements on our subject's preferences. It should go without saying, though, that any resulting 'fuzziness' with respect to U will ramify up the system and result in an even less specific credence function.

5 Conclusion

In Section 2.1, I said I wanted a representation with axioms that ordinary subjects come close to satisfying relative to some interesting domain under minimally idealised conditions, with an intuitively plausible decision rule $\mathscr{R}$ and relatively few constraints $< R_1, R_2, \ldots, R_n >$ on Cr and U. Furthermore, the uniqueness conditions on Cr and U should be reasonably restrictive. To conclude, then, I want to offer a brief evaluation of the foregoing results in light of these goals.

Let's begin with $< R_1, R_2, \ldots, R_n >$. In establishing the existence and uniqueness conditions for Theorem 3's Cr, we have only officially had to assume that it is real-valued and defined on some prespecified set of propositions $\mathcal{P}$, which need not have an algebraic structure. Likewise, we have only needed to assume U is

[17] Taking this strategy also requires a slight re-interpretation of $\succ$ and $\sim$, as they were defined in Section 3.4. We can keep the definitions of $\succ$ and $\sim$ in terms of $\succsim$, but we should only say that S strictly prefers P to Q (or is indifferent between them) if $P \succ Q$ (or $P \sim Q$) on all coherent completions of $\succsim$.

real-valued, and defined on a prespecified space of outcomes $\mathcal{O}$.[18] Any other properties of Cr and U (e.g., that $Cr(P) = 1 - Cr(\neg P)$, that $Cr(P) \in [0, 1]$, that $Cr(\pi) = \frac{1}{2}$, etc.) are all then consequences of the axioms, not independent restrictions we need to place upon the functions for the purposes of proving the representation and uniqueness results.

Of course, within those axioms are a number of non-trivial structural constraints on $\mathcal{P}$ and $\mathcal{O}$. From S2 and A1.2, $\mathcal{P}$ has to be closed under negation. Also from S2 and A1.2, for every complementary pair of propositions P and $\neg P$ in $\mathcal{P}$, there has to be a pair of outcomes o_1 and o_2 of unequal value bearing the $\rightharpoonup$ relation to P and $\neg P$ respectively. From S2 and A1.1, for every outcome o, there is at least one proposition $\pi \in \Pi$ such that $Cr(\pi) = \frac{1}{2}$ and $o' \rightharpoonup \pi$ for some outcome o' equal in value to o. And from S3, for every gamble that can be found in $\mathcal{G}$, certain conjunctions need to exist in $\mathcal{O}$. (Whether this automatically makes $\mathcal{O}$ infinite or not depends on how fine-grained you like your propositions). Each of these conditions will still need to be met even if we pursue any of the weakened versions of the theorem discussed in Section 4, but I'm not inclined to see any as especially problematic.

The fact that $\mathcal{O}$ and $\mathcal{P}$ have to be specified independently presents an interesting challenge, and one which is far too often overlooked. That is, we need to have some minimal access to what propositions our subject has credences and utilities towards in order to flesh out the purely formal constructions that we have been employing to talk about her preferences. At best, Theorem 3 gives us a way of pinning down the particular *degrees* of confidence and desire that S attaches to different propositions, *given* a specification of just what propositions she has credences and utilities towards. What Theorem 3 doesn't do is give us a way of getting at what sorts of propositions she has these attitudes towards in the first place.

But this limitation is not had by Theorem 3 alone. As noted in Section 2.1, Savage requires us to stipulate at the outset the domains of both Cr and U. On the other hand, Jeffrey makes the assumption that every proposition towards which S has preferences is also one towards which S has credences. He then also assumes that S has preferences towards a non-specific bottomless algebra of propositions. Neither assumption is trivial, and I would prefer to avoid making either in the absence of strong reasons in their favour.

The flexibility that my $\mathcal{O}$ and $\mathcal{P}$ have therefore seems like a distinct advantage, and the fact that their domains cannot be derived from the facts about preference is not obviously a disadvantage. It is not at all clear that it's possible to determine an appropriate domain for Cr merely given S's preferences.[19] It's plausible that we can derive an appropriate domain for U from S's preferences, for obvious reasons.

[18] The extension of U to $\mathcal{O} \cup \mathcal{G}$ is an optional extra. It's straightforward to restate the theorem such that U is only defined for outcomes, with a distinct function EU on $\mathcal{G}$ characterised in terms of Cr and U.

[19] In connection with this point, it's worth pointing out that Savage's credence functions are fundamentally incapable of representing subjects' credences regarding their own actions and anything probabilistically dependent upon them [14]. The same applies to every theorem based on a similar kind of formal framework. If agents do have credences towards the relevant kinds of proposition, then no Savagean theorem will let us fully pin down all of the credence facts using information from preferences alone.

The real difficulty is in getting a specification of $\mathcal{P}$ out of $\succsim$. And it's certainly not unreasonable to think that we may *not* be able to extract a complete specification of *Cr*'s domain solely from *S*'s preferences. If this is so, then the need to treat $\mathcal{O}$ and $\mathcal{P}$ as primitives rather than trying to derive them from $\succsim$ just reflects the facts about what kinds of information we can expect to get out of preferences alone.

Regarding whether ordinary agents generally satisfy my axioms and whether the basic expected utility norm for gambles in $\mathcal{G}$ is plausible, I cannot say a great deal for certain. Ultimately, this is an empirical matter, not one that's suited for armchair speculation. I have argued that we can do without Theorem 3's least plausible axioms (A2.2, A8, and A9), at the cost of greater indeterminacy in *Cr* and *U*. Of these, axiom A9 in my system is the most similar to those axioms of more classical expected utility theorems which have created the most controversy; e.g., von Neumann & Morgenstern's 'Independence' axiom and Savage's P2 and P4.

Of those axioms that remain, I think it's plausible that ordinary agents come close to satisfying the axioms under fairly normal conditions; i.e., conditions where they have time to reflect, are not under the influence of drugs, anxiety and alcohol, and so on. I've already given reasons for thinking that A1.1 is generally satisfied, and it's certainly hard to imagine widespread failures of A3, A7 and A1.2. Although contestable, A2.1 seems on the whole to be empirically plausible [27, p. 13]. A4 is only really plausible once it has been translated in terms of $\geq^d$, which is problematic: the translation presupposes the adequacy of Definition 2 as a definition of $\geq^d$, which in turn presupposes the accuracy of an expected utility representation for $\succsim$ on $\mathcal{G}$. A5 and A6 are partially structural axioms with analogues shared by the large majority of decision-theoretic representation theorems, so the present theorems are at least no worse than their competitors in that respect.

Nevertheless, it will be worth saying a few more things about the empirical evidence relating to expected utility theory generally. The orthodox opinion is that ordinary agents generally do not satisfy the axioms associated with any of the classical theorems for expected utility theory; there is a long history of experiments pointing to violations with Savage's theorem in particular. However, these theorems aim for a much more general representation of subjects' preferences than I have aimed for: their 'relevant domain $\mathscr{D}$' is usually the set of all things towards which a subject does (or *might*) have preferences. Savage's axioms put restrictions on arbitrarily complex 'acts', the vast majority of which won't be representable in our set of simple two-outcome gambles $\mathcal{G}$. Consequently, some of the evidence that's relevant to expected utility theory *qua* general theory of decision-making is difficult to apply in the present context. For example, without three-outcome gambles it's not possible to formulate even the very simple set-up needed for the Allais paradox.

Classical expected utility theorems also typically aim towards the *probabilistic* representation of subjects' credences. For that reason, they require additional constraints and a more complex space of gambles which are capable of partitioning the possibilities into more than just a single 'win condition' and its negation. Ramsey's own representation theorem begins with axioms very similar to Theorem 2's, but he ensures probabilistic coherence only through the addition of further axioms for *n*-outcome gambles. See Bradley [4, esp. Section 2.3] for discussion on how this is done. Note that Bradley's proofs require stronger assumptions about the logical

structure of $\mathcal{P}$ than I have made, and the additional axioms (R10, R11) are only intuitively plausible to the extent that the subject is fully cognizant of that extra structure. Generally, something like this is always going to be needed if *Cr* is to be a probability function.

So, one lesson that we can take from the empirical literature seems to be this: we are not (probabilistically coherent) expected utility maximisers *in general*. And with our bounded capacities, we'll be less and less likely to satisfy the norms of expected utility theory as the options we deliberate on become more and more complex. However, for the simple cases, something like expected utility theory probably comes close to the truth a lot of the time. This much *is* supported by the evidence. Even psychological models that are explicitly designed to accommodate the empirical evidence of our deviations from expected utility theory typically bear a very close resemblance to that theory. With just a few exceptions, they involve a real-valued (though usually non-additive) *Cr* and a real-valued *U*, combined in something like expectational form, with the basic decision-making rule being that an agent will pick the option which has the highest *Cr*-weighted average utility. There may be some bells and whistles added in to deal with risk-attitudes and framing effects, non-additive or imprecise credences and so on, but—as a rule—contemporary models of decision-making closely resemble the basic model of expected utility theory.[20] It would be a mistake to infer, from the very large amount of evidence that we are not expected utility maximisers, that we are therefore nothing like expected utility maximisers—nor that we wouldn't be under some fairly minimal idealisations.

The axioms put forward here don't say anything about what our subject's preferences have to be like for the kinds of arbitrarily complicated (and often imaginary) choices that Savage's framework lets us model; but that is not their point. Instead, the very simple gambles in $\mathcal{G}$ provide a limited domain for $\succsim$ wherein something like expected utility theory is more likely to be accurate, and (hence) where an ordinary agent's credences and utilities will most clearly shine through in her preferences. This can be true even for the frequently irrational—so long as they maximise expected utility with respect to $\mathcal{G}$, we can show that there's enough information contained in their preferences to play a significant role in pinning down what their credences and utilities have to be.

Acknowledgments I would like to thank an anonymous referee for this journal for detailed and helpful comments. I am grateful to Ben Blumson, Rachael Briggs, David Chalmers, Daniel Elstein, Jessica Isserow, Al Hájek, James Joyce, and Robbie Williams, for helpful comments and discussion on the paper and its immediate predecessors. Thanks also to audiences at the ANU, the University of Leeds, the National University of Singapore, and the LSE. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 312938.

[20]This is essentially the case, for example, of Tversky and Kahneman's [40] cumulative prospect theory, widely thought to be the most empirically accurate model of decision-making so far developed. Simplifying somewhat, CPT models agents as preferring acts with the greatest μ-weighted average utility, where μ is a monotonic function from a set of events to [0, 1]. The 'decision-weight' μ is usually taken to be decomposable into the subject's credences and her attitude towards risk (cf. [43]).

References

1. Alon, S., & Schmeidler, D. (2014). Purely subjective maxmin expected utility. *Journal of Economic Theory*, *152*, 382–412.
2. Anscombe, F.J., & Aumann, R.J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics*, *34*(2), 199–205.
3. Aumann, R.J. (1962). Utility theory without the completeness axiom. *Econometrica*, *30*(3), 445–462.
4. Bradley, R. (2001). Ramsey and the measurement of belief. In Corfield, D., & Williamson, J. (Eds.) *Foundations of Bayesianism* (pp. 261–275). Kluwer Academic Publishers.
5. Buchak, L. (2023). *Risk and rationality*. Oxford: Oxford University Press.
6. Chalmers, D. (2011a). Frege's puzzle and the objects of credence. *Mind*, *120*(479), 587–635.
7. Chalmers, D. (2011b). *The nature of epistemic space*, (pp. 60–107). Oxford: Oxford University Press.
8. Cozic, M., & Hill, B. (2015). Representation theorems and the semantics of decision-theoretic concepts. *Journal of Economic Methodology*, *22*, 292–311.
9. Davidson, D. (1980). Toward a unified theory of meaning and action. *Grazer Philosophische Studien*, *11*, 1–12.
10. Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, *87*(6), 279–328.
11. Dogramaci, S. (forthcoming). Knowing our degrees of belief. *Episteme*.
12. Easwaran, K. (2014). Decision theory without representation theorems. *Philosopher's Imprint*, *14*(27), 1–30.
13. Eells, E. (1982). *Rational decision and causality*. Cambridge: Cambridge University Press.
14. Elliott, E. (forthcoming a). Probabilism, representation theorems, and whether deliberation crowds out prediction. *Erkenntnis*.
15. Elliott, E. (forthcoming b). Ramsey without ethical neutrality: a new representation theorem. *Mind*.
16. Fishburn, P.C. (1981). Subjective expected utility: a review of normative theories. *Theory and Decision*, *13*, 139–199.
17. Gilboa, A.I., Postlewaite, A., & Schmeidler, D. (2012). Rationality of belief or: why savage's axioms are neither necessary nor sufficient for rationality. *Synthese*, *187*, 11–31.
18. Harsanyi, J. (1977). On the rationale of the Bayesian approach: comments of Professor Watkins's paper. In Butts, R.E., & Hintikka, J. (Eds.) *Foundational Problems in the Special Sciences* (pp. 381–392). Dordrecht: J. Reidel.
19. Jeffrey, R. (1986). Bayesianism with a human face. *Minnesota Studies in the Philosophy of Science*, *10*, 133–156.
20. Jeffrey, R.C. (1968). Probable knowledge. *Studies in Logic and the Foundations of Mathematics*, *51*, 166–190.
21. Jeffrey, R.C. (1990). *The logic of decision*. Chicago: University of Chicago Press.
22. Joyce, J. (2015). The value of truth: a reply to Howson. *Analysis*, *75*, 413–424.
23. Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*, 263–291.
24. Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: additive and polynomial representations* Vol. I. Academic Press.
25. Levin, I.P., Schneider, S.L., & Gaeth, G.J. (1998). All frames are not equal: a typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*, 149–188.
26. Lewis, D. (1974). Radical interpretation. *Synthese*, *27*(3), 331–344.
27. Luce, R.D. (1992). Where does subjective expected utility fail descriptively? *Journal of Risk and Uncertainty*, *5*, 5–27.
28. Luce, R.D., & Krantz, D.H. (1971). Conditional expected utility. *Econometrica*, *39*(2), 253–271.
29. Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.
30. Maher, P. (1997). Depragmatized dutch book arguments. *Philosophy of Science*, *64*(2), 291–305.
31. Meacham, C.J.G., & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, *89*(4), 641–663. doi:10.1080/00048402.2010.510529.
32. Pettit, P. (1991). Decision theory and folk psychology. In Bacharach, M., & Hurley, S. (Eds.) *Foundations of Decision Theory: Issues and Advances* (pp. 147–175). Oxford: Basil Blackwater.
33. Rabinowicz, W. (2012). Value relations revisited. *Economics and Philosophy*, *28*, 133–164.
34. Ramsey, F.P. (1931). Truth and probability. In Braithwaite, R. B. (Ed.) *The foundations of mathematics and other logical essays* (pp. 156–198). London: Routledge.
35. Rinard, S. (2015). A decision theory for imprecise credences. *Philosopher's Imprint*, *15*, 1–16.

36. Savage, L.J. (1954). *The foundations of statistics*. New York: Dover.
37. Schervish, M.J., Seidenfeld, T., & Kadane, J.B. (1990). State-dependent utilities. *Journal of the American Statistical Association*, *85*, 840–847.
38. Seidenfeld, T. (2004). A contrast between two decision rules for use with (convex) sets of probabilities: gamma-maximin versus e-admissibility. *International Journal of Approximate Reasoning*, *140*, 69–88.
39. Troffaes, M.C.M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29.
40. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.
41. Van Schie, E.C.M., & Van Der Pligt, J. (1995). Influencing risk preference in decision making: the effects of framing and salience. *Organizational Behavior and Human Decision Processes*, *63*, 264–275.
42. von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
43. Wakker, P.P. (2004). On the composition of risk preference and belief. *Psychological Review*, *111*, 236–241.
44. Walley, P. (1999). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, *24*, 125–148.
45. Weirich, P. (2004). *Realistic decision theory: rules for nonideal agents in nonideal circumstances*. Oxford: Oxford University Press.
46. Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, *67*(1), 45–69.